Computational Analysis of Historical Narratives through Large Language Models

Isuri Anuradha Nanomi Arachchige, BEng (Hons)

School of Computing and Communications

Lancaster University

A thesis submitted for the degree of

Doctor of Philosophy

June, 2026

Computational Analysis of Historical Narratives through Large Language Models

Isuri Anuradha Nanomi Arachchige, BEng (Hons).

School of Computing and Communications, Lancaster University

A thesis submitted for the degree of *Doctor of Philosophy.* June, 2026.

Abstract

Over the last decade, Large Language Models (LLMs) have been pushing the boundaries of artificial intelligence in creativity, language generation, and specialised problem-solving. Their ability to understand and generate human-like text makes them particularly useful for different Natural Language Processing (NLP) tasks such as information extraction, text classification, summarisation, and question answering. With these advancements, the integration of LLMs into digital humanities for sensitive, domain-specific contexts remains largely unexplored. This is mainly due to the unstructured, context-dependent nature of historical texts such as oral Holocaust narratives, which present unique linguistic and ethical challenges. This study focuses on developing domain-specific NLP techniques for processing oral narratives within the broader field of digital humanities. The first phase of the thesis introduces a domain-specific framework for extracting named entities and relationships from historically and culturally sensitive narratives by employing state-of-the-art information extraction techniques. Second, we propose a novel, lightweight, and reproducible adapter-based architecture for information retrieval from oral narratives, which integrates advanced retrieval-augmented generation (RAG) techniques. Third, we construct a knowledge graph to systematically capture and analyse common patterns and insights across the narratives. Given the rapid emergence of LLMs and their increasing application in sensitive historical domains such as Holocaust research, this study critically analyses the ethical challenges associated with using LLMs in historically sensitive research. Overall, the research is designed with a flexible and modular architecture, enabling reproducibility and extension to similar historical documents that have yet to be digitised in archival collections.

Acknowledgements

This research journey has been long and challenging, both intellectually and personally. Choosing to work at the intersection of computer science and the history of the Second World War required not only intellectual courage but also a willingness to carry the emotional weight of the material every day. That I have been able to complete this work is entirely due to the generosity, guidance, and humanity of the people named here.

I owe my deepest gratitude to Professor Ruslan Mitkov, who believed in this research from the very beginning at the University of Wolverhampton and whose support never wavered throughout the entire journey. He has been far more than a supervisor—he has been a mentor, an encourager, and an academic compass at every stage of this work. He treated me with the warmth and affection of a daughter by staying by my side always, and I am profoundly grateful for the trust he placed in me and for the wisdom he shared so generously.

I would like to express my sincere thanks to Professor Paul Rayson, who supervised me through a significant period of this journey and whose academic guidance advanced my research in ways I could not have achieved alone. Changing universities in the middle of a PhD is not a simple undertaking, and his mentorship made that transition possible.

Several key people supported me from the very start and shaped the interdisciplinary character of this work in ways that proved essential. Dr Le An Ha of the University of Wolverhampton provided consistent support and guidance that grounded the computational side of this research. Professor Dieter Steinert of the History Department of the University of Wolverhampton brought a depth of historical expertise that ensured this work remained anchored in scholarly rigour as well as technical innovation. I would also like to thank Dr Ingo Frommholz, who supervised and supported me during part of my time at the University of Wolverhampton. His guidance during that period contributed meaningfully to the development of this research. Working across the boundary between computer science and history is a challenge that these two disciplines rarely invite, and their willingness to bridge that divide made this thesis possible. I am also grateful to Dr Tharindu Ranasinghe, whose support and encouragement helped me take the very first steps toward beginning this PhD. The journey starts long before the first chapter is written, and his role in that beginning was more significant than he may realise. I must pay tribute to Dr Vinita Nahar, one of my supervisors in the early stages of this PhD, who passed away suddenly and far too soon. She was a wonderful person whose kindness, enthusiasm, and belief in this research left a lasting impression on me. I carry her memory with me, and I hope this completed thesis honours the encouragement she gave so wholeheartedly.

Every PhD student needs at least one person who believes in them unconditionally and without

reservation, basically someone who is simply there. For me, that person has been my brother, not by blood, Damith Premasiri. We laughed together, we faced difficult moments together, and we grew through it all. A PhD is sustained not only by academic support but also by the people who make life outside research feel human and hopeful. I am grateful to Emma Franklin, Suman Hira, Hansi Hettiarachchi, Alistair Plum and the whole Research Group of Computational Linguistics at the University of Wolverhampton for the happy memories and the kindness they offered at various points along the way. Each of you made this journey lighter than it would otherwise have been. I would like to extend my thanks to the colleagues and friends at Lancaster University whose companionship made the day-to-day of research life a pleasure: Ignatius Ezeani, Saad Ezzini, Daisy Lal and Mo El-Haj. Thank you for the wonderful times and the warmth you brought to that period of the journey.

I would also like to express my gratitude to the European Holocaust Research Infrastructure (EHRI) and CLARIN EU for their invaluable institutional support and for providing access to the resources and networks that underpinned this research. In particular, I would like to thank Rachel Pistol from EHRI and Martin Wynne and Francesca Frontini from CLARIN EU, whose engagement with this work went beyond institutional formality and into genuine scholarly collaboration. I also want to acknowledge, with honesty and affection, that living with me is not always an easy thing. I am someone who is perpetually absorbed in research, frequently travelling, and not always the most patient of people. Through all of that, Nadeesha, my fiancé, has stood beside me with tolerance, steadiness, love, and trust. He gave me everything I needed, often before I knew I needed it, and the completion of this thesis owes more to his quiet, unwavering support than these few words can properly reflect.

Finally, and most deeply, I want to thank my mother and father. As their only child, leaving home was not a simple decision for any of us. Their willingness to support me, to trust me, and to believe in a path they could not always fully see has been the foundation beneath everything I have built during these years. I hope this thesis makes you proud, as you have always made me feel that whatever I reached for was within my grasp. This thesis belongs to you as much as it belongs to me. I would also like to thank my uncle, Loku Bappa, whose presence and warmth have always meant a great deal to me, even across the distance. To my best friends- Miuru Wijemanna, Madushani, Vishka, Ruvan Weerasinghe, Rajitha, Bimsara and all other friends proved distance has never diminished what we share. No matter how far apart we have been, a single message has always been enough to bring you to my side without a moment's hesitation. To everyone named here and to anyone whose kindness I have been fortunate enough to receive along the way: thank you.

List of Publications

Contributing publications

Isuri Anuradha and Martin Wynne (2026). "Proceedings of the Second Workshop on Holocaust Testimonies as Language Resources (HTRes)@ LREC 2026". In: *Proceedings of the Second Workshop on Holocaust Testimonies as Language Resources (HTRes)@ LREC 2026*

Isuri Anuradha, Tharindu Ranasinghe, Ruslan Mitkov, and Ingo Frommolz (2026). "LiLADH: An Open Retrieval Resource for Digital Humanities Archival Corpora". In: *Submitted to the Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM 2026)*. Under review (**Chapter 7**)

Isuri Anuradha, Deshan Koshala Sumanathilaka, Ruslan Mitkov, and Paul Rayson (2025). "Toponym Resolution: Will prompt engineering change expectations?" In: *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing-Natural Language Processing in the Generative AI Era*, pp. 95–104 (**Chapter 4**)

Isuri Anuradha, Ruslan Mitkov, et al. (2025). "HoloBERT: Pre-Trained Transformer Model for Historical Narratives". In: *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing-Natural Language Processing in the Generative AI Era*, pp. 105–110 (**Chapter 4**)

Isuri Anuradha, Martin Wynne, Francesca Frontini, and Alistair Plum (2024). "Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes)@ LREC-COLING 2024". In: *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes)@ LREC-COLING 2024*

Isuri Anuradha, Ruslan Mitkov, Vinita Nahar, et al. (2023). "Evaluating of Large Language Models in Relationship Extraction from Unstructured Data: Empirical Study from Holocaust Testimonies". In: *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pp. 117–123 (**Chapter 5**)

Isuri Anuradha, Le Ha, Ruslan Mitkov, and Johannes-Dieter Steinert (2023). "Enhancing named entity recognition for holocaust testimonies through pseudo labelling and transformer-based models". In: *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing*, pp. 85–90 (**Chapter 4**)

Additional publications

Tharindu Ranasinghe, **Isuri Anuradha**, Damith Premasiri, Kanishka Silva, Hansi Hettiarachchi, Lasitha Uyangodage, and Marcos Zampieri (2025). "Sold: Sinhala offensive language dataset". In: *Language Resources and Evaluation* 59.1, pp. 297–337

Tharindu Ranasinghe, Hansi Hettiarachchi, Nadeesha Chathurangi Naradde Vidana Pathirana, Damith Premasiri, Lasitha Uyangodage, **Isuri Anuradha**, Alistair Plum, Paul Rayson, and Ruslan Mitkov (2025). "Sinhala encoder-only language models and evaluation". In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8623–8636

Isuri Anuradha, Francesca Frontini, Ruslan Mitkov, and Paul Rayson (2025). "Proceedings of the First on Natural Language Processing and Language Models for Digital Humanities". In: *Proceedings of the First on Natural Language Processing and Language Models for Digital Humanities*

Ruvan Weerasinghe, **Isuri Anuradha**, and Deshan Sumanathilaka (2025). "Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages". In: *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Those who cannot learn from history are doomed to repeat it.

George Santayana (Survivor)

This chapter initiates the thesis by providing a broad context for its motivation, research questions and contributions. Over the past few decades, humanity has always been torn between its capacity for connection and its propensity for conflict. This tension has re-emerged with the advancement of technology. Digital preservation serves to safeguard the memories of those silenced by violence and genocide, while artificial intelligence provides innovative pathways for education and remembrance. However, these same technologies also carry the potential to amplify distortion, corrupt historical truth, and propagate hatred across societies. This thesis investigates how we can preserve historical integrity in digital spaces by combining humanity, historical memory, and technological advancement.

1.1 Background

Today, we inhabit an era of digital humanity: human existence is increasingly mediated, documented, and experienced through digital technologies. Most of our social interactions occur across virtual platforms worldwide, and our memories are stored in servers which are accessible through search engines within less than a

minute. This digitalisation has created new possibilities for connection, education, and preservation, enabling democratising access to information and offering digital tools to make human life more comfortable. However, digital humanity also presents profound challenges. The same technologies that preserve history can distort it, and the artificial intelligence that enhances learning can spread misinformation. For instance, large language models have been shown to hallucinate biographical details when queried about Holocaust survivors, generating plausible but entirely fabricated accounts of names, dates, and locations Digital Watch Observatory, 2024. This duality demands careful, principled governance of how digital tools are designed and deployed, particularly when the materials they engage with carry historical and moral weight. Digital humanity confronts a particularly acute responsibility when addressing documents related to wars and conflicts. Unlike other historical materials, war documentation carries the weight of trauma, loss, and collective memory. These documents, which range from testimonies of genocide survivors and prisoner accounts to battlefield letters, represent not merely historical data but an inviolable trust. Furthermore, the digitisation of conflict-related materials presents unique challenges in preserving the dignity of victims while making their stories accessible, contextualising atrocity without sanitising its horror and protecting sensitive information while serving the public's right to historical truth.

Traditional methods of processing these collections have proven inadequate to the task. Manual cataloguing and analysis, while respectful and nuanced, cannot keep pace with the volume of materials requiring preservation. A single historical event may produce hundreds of thousands of documents, which require years or decades to properly index and cross-reference by human archivists. This process has led to consequences, where witnesses age and pass away before their experiences are integrated into the historical record, perpetrators exploit the slow pace of documentation to bypass accountability, and communities lose access to materials that could support healing and reconciliation efforts. Additionally, manual processing reinforces existing inequalities in whose stories are prioritised, as archives

with more capital and funding receive attention, while other materials deteriorate in condition. The linguistic barriers present in these documents are equally challenging, as documents in low-resource languages with regional dialects remain untranslated and inaccessible for international researchers and audiences. Moreover, humans in this process could carry their own trauma burden, as archivists and researchers experience traumatisation from prolonged exposure to accounts of violence and suffering.

The Holocaust stands as one of history's darkest chapters in humanity, where a systematic genocide perpetrated by Nazi Germany and its collaborators resulted in the murder of six million Jews, along with millions of others, such as Roma, Gipsies, people with disabilities, political dissidents, homosexuals, and Jehovah's Witnesses. From 1933 to 1945, the Nazi regime transformed antisemitic prejudice into a state-sponsored industrial-scale mass murder, establishing concentration camps and implementing policies designed to eliminate entire Jewish populations. As the generation of Holocaust survivors ages and passes away, it is important to preserve their stories for future generations, not just as historical facts, but also the human experiences behind the statistics. Survivor testimonies serve as irreplaceable historical documents and moral imperatives that bring us the depths of human cruelty and the resilience of the human spirit.

Oral Holocaust testimonies represent first-hand accounts that convey direct and personal perspectives on the experiences of survivors, victims, and perpetrators. These testimonies hold immense historical and cultural value, offering insights essential for teaching the human dimensions of history that standard official records alone cannot convey. The collection of Holocaust survivor testimonies began almost immediately after liberation. In displaced persons' camps across Europe, survivors felt an urgent need to document what had happened to honour the dead and feared that their experiences might be forgotten or denied. Early efforts were initiated by interviewing survivors and recording their stories in journals and audio recordings. After the latter half of the twentieth century, the systematic collection of oral

histories gained momentum, and institutes such as Yad Vashem[1], the United States Holocaust Memorial Museum[2], and the Visual History Foundation[3] were established to create formal repositories for survivor testimonies. While written documents, photographs, and physical evidence provide crucial historical documentation, oral testimonies provide real experiences about the Holocaust. Oral testimonies capture the emotional language of what survivors have experienced, from the hesitations, silences, sudden floods of memory, tears, and moments of unexpected warmth or humour that punctuate even the darkest narratives.

In Holocaust testimonies, survivors describe not only the major events, such as deportations, selections, forced labour, and liberation, but also the small details that reveal what it meant to live through such horror. These accounts also reveal the diversity of Holocaust experiences. Each testimony is unique, reflecting diverse geographical locations, varying types of persecution, diverse ages, and varied circumstances. Some survivors hid in attics and forests, while others were directly transported to concentration camps or death camps, some survived ghettos, and others escaped through various means. Some were children, young adults or parents. These voices remind us of the catastrophe that destroyed millions of individual lives, each with its own story. The scale of testimony collections far exceeds what manual cataloguing can process within meaningful timeframes. Linguistic barriers leave large portions of multilingual archives difficult to access, limiting their use in research and education. At the same time, many witnesses are ageing and passing away before their accounts can be fully integrated into the historical record. Archivists working with these materials also experience significant emotional strain due to prolonged exposure to testimonies describing traumatic events. Taken together, these challenges highlight the need for computational approaches that can operate at the scale, speed, and analytical depth required by large testimony collections. Natural Language Processing (NLP) offers a promising means to support this work

[1]https://www.yadvashem.org/

[2]https://www.ushmm.org/

[3]https://vha.usc.edu/home

by enabling the automated processing, organisation, and analysis of large volumes of textual testimony.

1.2 Natural language Processing & Language Models

Natural Language Processing (NLP) enables machines to process, understand, and interpret human language in structured and machine-readable formats (Torfi et al., 2021). By combining linguistics and computer science, NLP bridges the gap in understanding the ambiguous, context-dependent, and nuanced nature of human communication. NLP has become popular within the different domains, including digital humanities, where scholars seek computational methods to analyse, preserve, and generate new insights from vast textual corpora spanning centuries of human expression. Traditional NLP methods relied on handcrafted rules and statistical models that were able to capture grammatical structures, semantic relationships, and contextual patterns (Torfi et al., 2021). However, when these rule-based models were applied to domain-specific, predefined tasks, they struggled with the inherent variability and creativity of natural language (Lauriola, Lavelli, and Aiolli, 2022). Machine learning-based statistical approaches outperformed rule-based models in various tasks, including translation, sentiment analysis, named entity recognition, and information extraction. Nevertheless, due to their dependence on annotated corpora and feature engineering, as well as their inability to capture the deeper semantic understanding, these models remained limited.

The emergence of large language models (LLMs) has revolutionised both NLP as a technical field and its applications within digital humanities scholarship (J. Liu et al., 2024). Built upon transformer architecture, LLMs are trained on massive corpora of text data, encompassing hundreds of billions or even trillions of words drawn from books, articles, websites, and digitised historical documents (Mienye et al., 2025). This scale of training enables performing remarkably across diverse

domains and tasks, including natural language understanding, generation, complex reasoning, and creative composition. Unlike traditional approaches, LLMs perform well on tasks that they were not explicitly trained on initially, adapt to new domains with minimal examples, and generate human-quality text, maintaining coherence across extended documents. For digital humanities scholars, LLMs could act as powerful analytical tools by analysing literary style, tracing the evolution of concepts across historical periods, identifying intertextual relationships within vast archives, and uncovering patterns that are impossible to detect through traditional close reading alone (J. Liu et al., 2024). However, integrating LLMs into digital humanities practice raises methodological and ethical questions that require careful consideration. Since most LLMs are primarily trained on contemporary English-language, web-crawled text and applied to historically specific materials, several common challenges arise, such as issues of bias, representation, and interpretive authority. Given that, the integration of NLP into digital humanities is a complex process that extends beyond simply applying computational tools to a general textual corpus.

1.3 Motivation

This thesis develops an understanding of a profound ethical and historical imperative: the urgent responsibility to listen to, preserve, and comprehend the stories of Holocaust survivors while the opportunity still exists. We stand at a critical time in history, transitioning from a living memory sustained by survivors to a post-testimonial era where their direct voices will fall silent. The hundreds and thousands of oral testimony collections present an irreplaceable corpus of human experience. These are not merely historical records; they are the embodied memories of individuals who endured the unimaginable, providing an opportunity to comprehensively study the stories of survivors and victims of a tragic event in human history. One of the motivations of this research is to honour these narratives

by ensuring that their full depth and complexity are accessible, moving beyond a paradigm where these archives are seen primarily as repositories for selective, manual consultation. The ambition is to transform them from static digital libraries into dynamic, computationally accessible resources that can answer complex, large-scale historical and sociological questions.

The second motivation of this thesis arises directly from the dual challenges of scale and complexity that these narratives present. Holocaust testimony is a genre of profound complexity. The testimonies are unstructured, fragmented by trauma, and rich in emotional and sensory details. Traditional archival research is inherently constrained when analysing and transcribing these tens of thousands of hours of video testimonies. This manual approach is not only slow but cognitively demanding, and the data itself presents semantic and technical challenges that existing computational tools cannot adequately address. These challenges include fragmented narratives, multilingual expression, and implicit references, which render existing computational approaches that adapt standard NLP toolkits insufficient for the task (Picheny, Zóltan Tüske, et al., 2019; Gref et al., 2022). Simple keyword searches and conventional topic modelling, commonly applied to cleaner textual data, similarly fail to capture the narrative depth, historical context, and traumatic experiences embedded in testimony collections (Williams et al., 2024). They risk reducing profound human experiences to sterile data points. For example, on some occasions, the survivor spoke about `Auschwitz`, but these models were unable to distinguish between a description of arrival, the daily struggle for survival, and the moment of liberation. This highlights the limitations of existing computational and language processing approaches in addressing the complexities of historical documents.

On motivation, this thesis is firstly inspired by the unprecedented opportunity to comprehensively study one of the stories of survivors and victims of a tragic period in human history. Secondly, it is motivated by the limitations of existing computational and language processing approaches in addressing the complexities of historical

documents, as well as the semantic and technical challenges inherent in spoken language. These challenges include fragmented narratives, multilingual expression, and implicit references that often hinder manual or traditional archival analysis, a process that can consume a prohibitive amount of time. By applying advanced NLP techniques, it is able to avoid information loss and identify clusters of survivors who shared similar experiences, construct relational networks linking individuals to specific deportation events, and trace post-war trajectories to understand patterns of displacement, resettlement, and community rebuilding. And finally, the ambition of the common good and contribution to human society through the transformation of novel language processing approaches into technology platforms that deliver insights for the betterment of humanity and preserve knowledge for future generations.

1.4 Research questions

Based on the motivation for the domain, the following research questions will be addressed in this thesis.

1. How can NLP techniques, including domain-aware entity recognition, prompt-based relationship extraction, and knowledge graph construction, be combined into a coherent framework for extracting and representing structured knowledge from Holocaust oral narratives? (Addressed by Chapters 4, 5, and 6)

2. What are the linguistic and structural challenges of Holocaust oral testimonies as unstructured data, and what corpus design and annotation strategies are required to make them computationally processable? (Addressed by Chapter 3)

3. How can a lightweight, parameter-efficient adapter architecture improve retrieval accuracy in RAG pipelines for domain-specific historical corpora without requiring large-scale model retraining? (Addressed by Chapter 7)

4. What ethical considerations arise when deploying large language models to process sensitive oral historical narratives, and how should these shape the design of NLP systems in digital humanities contexts? (Addressed by Chapter 8)

1.5 Research Contributions

This thesis makes four original contributions to the fields of natural language processing and digital humanities. As illustrated in Figure 1.1, the contributions have been arranged as below:

Contribution 1: A curated and annotated Holocaust oral testimony corpus

The foundation of the framework presented in Figure 1.1 is the data layer. This thesis presents a systematically curated and manually annotated corpus of Holocaust oral testimonies, constructed to address the unique computational challenges. Holocaust testimonies exhibit a distinctive set of linguistic properties that does not appear in general-purpose NLP training corpora. To make these testimonies amenable to computational processing, this research develops a manually curated annotation schema covering entities, relationships, and contextual markers relevant to Holocaust history. The corpus, its curation methodology, and the annotation guidelines constitute the gold dataset for all subsequent technical work in the thesis.

Contribution 2: A Domain-Aware knowledge extraction and representation framework

According to Figure 1.1, the information extraction pipeline represents the thesis's primary technical contribution: an end-to-end framework that transforms unstructured oral testimonies into structured, machine-readable knowledge. This framework comprises three tightly integrated components. First, a domain-aware named entity recognition (NER) model (Chapter 4) is developed to identify and contextualise historically specific entities such as persons, locations, organisations,

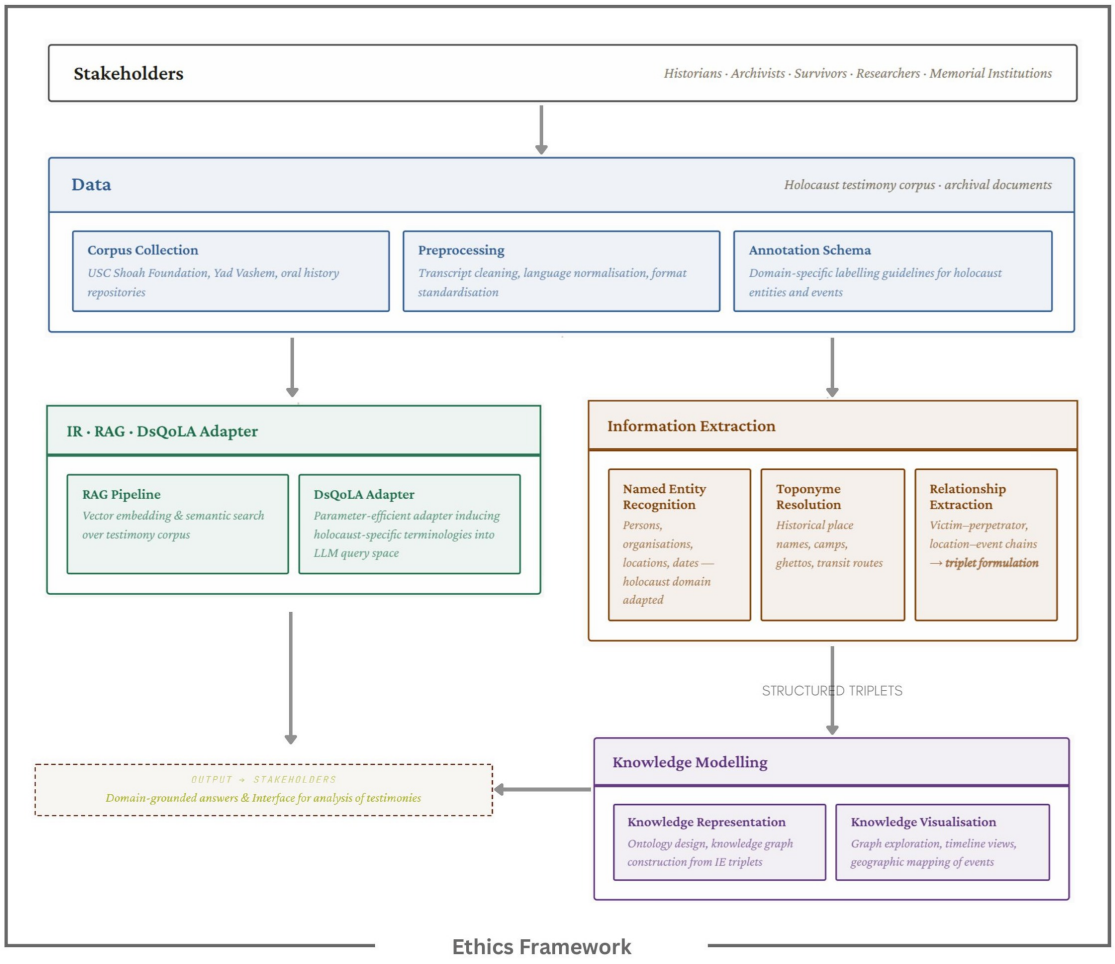

Figure 1.1: Overall Framework and Contribution Phases

and events. As an extended version of NER, separate experiments have been conducted on the toponym resolution. Second, a prompt engineering approach (Chapter 5) is introduced to extract domain-specific relationship types under low-resource conditions, producing structured triplets from narrative text with a minimal set of annotated examples. Third, these extracted entities and relationships are formalised into a knowledge graph (Chapter 6), designed for accessibility, usability, and societal relevance for stakeholders, including historians, archivists, and memorial institutions. As shown in Figure 1.1, the output of information extraction flows

directly into knowledge modelling via structured triplets, forming a coherent and traceable pipeline from raw testimony to navigable knowledge representation.

Contribution 3: DsQoLA - A lightweight domain-specific adapter for RAG pipelines

According to Figure 1.1 the Information Retrieval phase introduces DsQoLA (Domain-Specific Query and Language Adapter), a parameter-efficient adapter architecture designed to enhance retrieval accuracy in Retrieval-Augmented Generation pipelines operating over domain-specific historical corpora. General-purpose embedding models, trained predominantly on web-based text, fail to capture the specialised vocabulary and contextual semantics of Holocaust testimonies. The contribution demonstrates that meaningful domain adaptation is achievable through parameter-efficient methods, with broader applicability to other low-resource humanities corpora.

Contribution 4: An ethical framework for LLM processing of sensitive historical narratives

As illustrated in Figure 1.1, the ethical standards govern all the NER techniques which are used in this framework. This thesis identifies and systematises the ethical considerations that arise specifically when LLMs are applied to Holocaust testimonies and other sensitive oral historical corpora. Moving beyond general AI ethics discourse, this contribution articulates domain-specific responsibilities encompassing survivor dignity, the risk of historical misrepresentation and institutional accountability. The resulting ethical framework proposes concrete design principles for NLP systems operating in digital humanities contexts.

1.6 Thesis Outline

The remainder of the thesis is structured as follows.

Chapter 2- Background and Related Work: includes the review of relevant literature to address the interdisciplinary intersection of oral historical testimonies,

spoken linguistic patterns, and NLP techniques used for information extraction.

Chapter 3- Holocaust Testimonies as Oral Historical Corpora: presents the datasets related to the research and outlines the processes involved in their collection, curation, and annotation. It also discusses different types of oral narratives and the challenges of oral testimonies as unstructured data sources.

Chapter 4- Domain-Aware Entity Recognition from Oral Narratives: presents a domain-aware entity recognition approach tailored for extracting and contextualising information from historically significant events such as the Holocaust.

Chapter 5- Modelling Relationships in Historical Narratives: discusses an advanced prompt engineering approach that is able to extract domain-specific relationship types from the oral testimonies with a limited set of annotated data.

Chapter 6- Knowledge Extraction from Narrative Texts presents the framework implemented for knowledge extraction and representation through a knowledge graph, designed to ensure accessibility, usability, and relevance for broader societal use.

Chapter 7- Information Retrieval in Narrative Text Using a lightweight adapter: discusses developing a lightweight adapter that could plug into any of the retrieve-augmented generation (RAG) pipelines to extract domain-specific information.

Chapter 8- Ethics of LLMs in Processing Oral Historical Narratives: outlines the ethical concerns which are arising with the emergence of the LLMs in processing sensitive information such as Holocaust narratives.

Chapter 9- Conclusion and Future Work: presents the summary of the thesis and outlines potential directions for future research.

Chapter 2

Background and Related Work

For your benefit, learn from our tragedy. It is not a written law that the next victims must be Jews. It can also be other people. We saw it begin in Germany with Jews, but people from more than twenty other nations were also murdered.

Simon Wiesenthal (Survivor)

2.1 Chapter Introduction

This chapter provides a comprehensive review of the core concepts discussed in this research. It illustrates the most relevant studies and applications related to the computational analysis of historical and testimonial data, with a particular emphasis on Holocaust survivor narratives. Furthermore, it explores the interdisciplinary intersections between digital humanities, computational linguistics, and Holocaust studies. Where,

- Digital humanities: Digital humanities is an interdisciplinary field that applies computational methods and digital tools to traditional humanities research questions, enabling new forms of textual analysis, visualisation, and knowledge representation (Berry, 2012).
- Computational linguistics: Computational linguistics is the scientific study of language from a computational perspective, involving the development of

algorithms and models for processing, understanding, and generating human language (Schubert, 2020; Mitkov, 2022).

- Holocaust studies: Holocaust studies is an interdisciplinary academic field that examines the historical, cultural, and ethical dimensions of the Holocaust, integrating perspectives from history, literature, memory studies, and sociology (Stone, 2010).

This chapter highlights research gaps and methodological challenges in order to determine the direction and significance of the current study. It also critically assesses complex linguistic patterns and the multilingual nature of Holocaust testimonies and existing NLP applications and provides a systematic analysis of the limitations within current computational methodologies for this specific domain.

2.2 The Holocaust as a Historical Event

Throughout history, humanity has witnessed countless conflicts, wars, and persecutions driven by power, ideology, or prejudice. The Holocaust/*Shoah* was a catastrophic event in history, a systematic, state-sponsored persecution and genocide of six million European Jews by Nazi Germany during World War II (1941–1945) (United States Holocaust Memorial Museum, 2020). According to the historical sources, it began with the legal and social persecution of Jews following Adolf Hitler's rise to power in 1933 (Hilberg, 2003). Violence increased with events like Kristallnacht ('the Night of Broken Glass') in November 1938, a state-coordinated pogrom targeting Jewish homes, businesses, and synagogues (Bauer, 2002). The persecution became more radical with the outbreak of World War II in 1939. Jews were forced into ghettos, facing inhumane and overcrowded conditions.

The mass murder phase began in June 1941 with the invasion of the Soviet Union, when mobile killing units called the Einsatzgruppen murdered over 1.5 million people in mass shootings. The systematic plan to exterminate all European Jews, known as the `Final Solution`, was formalised at the Wannsee Conference in 1942. This

plan led to the establishment of death camps such as `Auschwitz-Birkenau`, where millions were killed in gas chambers (United States Holocaust Memorial Museum, 2020). While Jews were the main targets, the Nazis also persecuted and murdered millions of other groups, including Roma, Poles, and disabled people. As a part of this historical event, Holocaust testimonies stand as a crucial and irreplaceable source of knowledge and memory (United States Holocaust Memorial Museum, 2020). These first-hand accounts, recorded in different forms such as diaries, memoirs, or oral history interviews, provide an intimate and human perspective on the unimaginable horrors of the genocide. Holocaust testimonies capture the lived experiences of individuals, families torn apart, communities destroyed, and lives reduced to survival in the face of dehumanisation. Moreover, these sources discuss more than statistics and historical timelines, revealing individual experiences of fear, loss, resilience, and survival. Therefore, Holocaust testimonies have played a crucial role in shaping collective memory, countering denial, and educating future generations (United States Holocaust Memorial Museum, 2020).

2.2.1 Digitally transcribed Testimony Archives

Different types of documents related to the Holocaust, such as official records, photographs, and letters, and Holocaust testimonies have been digitised within the last decade (Brazzo and Speck, 2018). As a historical record, Holocaust testimonies ensure that the horrors committed during World War II are never forgotten. Moreover, testimonies can be considered educational resources which facilitate the understanding and awareness of the horrors that arise from prejudice, hatred, and systemic discrimination. Unlike other historical documents, which may provide a more detached or institutional perspective, testimonies capture individual experiences, memories, and reflections, making them invaluable for understanding the human impact of the Holocaust (Ionescu and Mitroiu, 2023).

This research mainly focuses on the Holocaust testimonies, which were collected in different archives and libraries throughout the world. These testimonies are

direct narratives from survivors, witnesses, and victims, offering deeply personal and emotional insights into the events they have endured in different languages. More information about those archives and libraries is included in Appendix A.

2.3 Language Use in Holocaust Oral Histories

Language enables effective communication between people. The use of language can vary significantly depending on the context of different situations (Halliday, 1989). Both spoken and written forms of language are shaped by the cultural, societal, and psychological circumstances of individuals' lives. Spoken and written languages have their unique complexities and challenges (Tannen, 1982). Reference to the oral histories related to the Holocaust: the intense emotional trauma endured by survivors is frequently expressed through their speech, representing the harshness of their experiences in ways that written narratives might not entirely reflect (Kraft, 2004). However, this influence of psychological trauma creates patterns of expression in the language, which require further linguistic exploration (Laub and Auerhahn, 2017).

In linguistic studies, language structure, form, and function are explained, including phonetics, syntax, semantics, and morphology (Gee, 2014). However, analysing the discourse to fully understand the language's structure, usage, and interpretation within various contexts is essential. Discourse analysis explains the context's deeper meaning beyond the literal words, revealing the interactions influenced by history, culture, politics, and identity (Fairclough, 2013). By applying discourse analysis to Holocaust narratives/testimonies, the structural format is clearly identified by the emotions and expressions conveyed through linguistic features, pauses, repetitions, and shifts in tone, offering a deeper understanding of their experiences and the sociocultural impact of the Holocaust as a historical event (Langer, 1993).

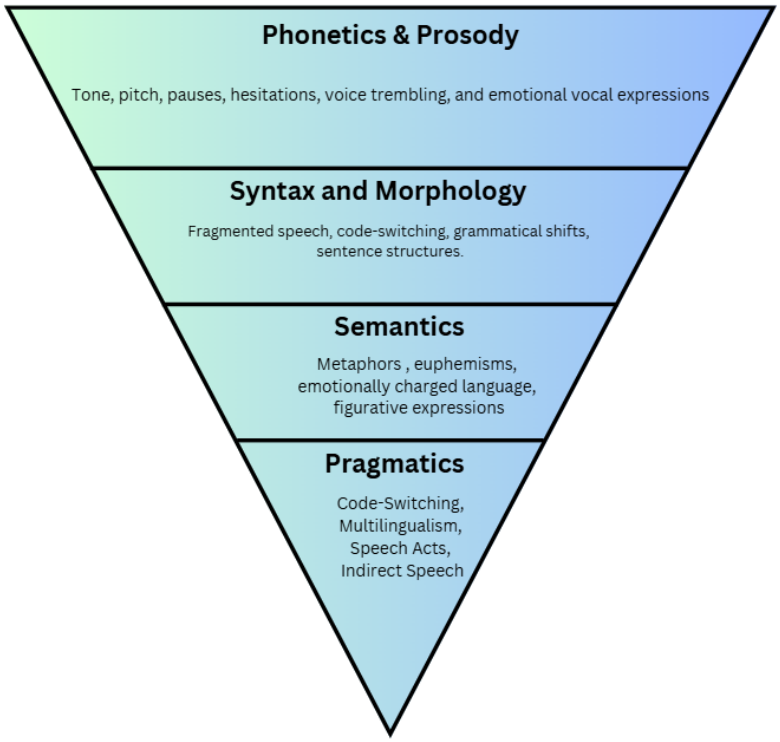

Figure 2.1: Linguistic feature Analysis of oral language

2.3.1 Spoken Language and Linguistic Features in Holocaust Oral Histories

Holocaust testimonies consist of different linguistic features, which require comprehensive analysis to examine the way the survivors reconstruct their experiences, convey trauma, and navigate memory through language (Schiffrin, 1994). Spoken language in Holocaust testimonies often presents unique challenges to understand due to linguistic variability intertwined with raw emotions and fragmented memories. When describing survivors' experiences, fragmented or disjointed sentences are common in the testimonies. This often occurs due to the trauma survivors endured during the Holocaust, which can disrupt the flow of memory, leading to pauses, hesitations, and incomplete thoughts. Additionally, the nature of these testimonies varies based on the survivor's age. Child survivors may have faded or incomplete memories due to their young age during the events. In contrast, adult survivors

Table 2.1: Linguistic features extracted from Holocaust Testimonies

Linguistic Features	**Samples from testimonies**
Neologisms	We carried out the 'Sonderaktion'—special action—and the 'material' was processed quickly. The Jews, they went to the 'shower rooms,' yes, and then it was finished. I... I gave orders, calm, always calm. *(Rudolf Höss, USHMM)*
Metaphor	The gas turned screams into silence, a harvest of death reaped in minutes. Order ruled the chaos; efficiency was our god. *(Rudolf Höss, USHMM)*
Collocation	We fought for a crust, a spoon, a rag—anything to keep the cold from biting. The Lager was a beast, and we were its prey. To live was to steal a day from death." *(Primo Levi, USHMM)*
Hyponym	We played violins, flutes, cellos—music to drown the screams. The SS sat like kings, but we were on the edge of a knife. In French, we whispered, 'C'est la fin du monde'—it's the end of the world. *(Fania Fénelon, USHMM)*
Idiom	I sent trains—freight cars, cattle cars—to the East. It was just paperwork, a cog in the wheel. The machine ran itself, and I kept my nose clean. *(Adolf Eichmann, Yad Vashem)*
Polysemy	The work broke us—every bone screamed for rest. We were shadows, not men, swallowed by the camp. I told my friend, 'Hold on, we'll rise again. (Primo Levi, USHMM)
Hyperbole	The gas took them all—my world ended in a second. I was a shell, hollowed out a million times over. I whispered to myself, 'Live, just live.' *(Halina Birenbaum, Yad Vashem)*

may still experience emotional distress and psychological trauma, affecting how they recall and describe their experiences of the Holocaust.

Specific terminology used in the Holocaust domain is a critical linguistic and powerful feature for conveying historical facts, emotional experiences, and cultural significance. Moreover, such terminology is frequently used to describe events before, during, and after the Holocaust, describing the efforts of survival and rebuilding in its aftermath. However, considering that terminology, Table 2.1 refers to different semantic features such as polysemy, hyponymy, metaphor and figurative language

that exist in the Holocaust testimonies (O'Donoghue, 2021). Polysemy is one of the key features that exist in the testimonies. It refers to the phenomenon where a single word carries multiple domain-specific meanings related to the Holocaust. Moreover, hyponymy is another linguistic feature which exists in the Holocaust testimonies. Most survivors use general or specific words when describing what they have witnessed during the Holocaust. Hyponymy is the association between specific and general terms, where a particular word (hyponym) falls under a broader category (hypernym) (Roca Lizarazu, 2017). Using hyponymy allows survivors to convey experiences with varying detail and emotional intensity. General terms may be used to describe broader experiences or when recalling events is too painful, while specific terms provide precise details that highlight the gravity of their suffering.

In Holocaust testimonies, metaphor and figurative language are the most important and frequently used linguistic elements. Holocaust survivors use linguistic features such as metaphors, similes, and symbolic expressions when direct language fails to capture or express how they have witnessed the trauma, suffering, resilience, and survival they have experienced (Bailey et al., 2020). Metaphors allow survivors to translate the inexpressible, painful memories into symbolic representations that may be easier to understand. Further, figurative language conveys the emotional intensity of incidents, locations, and events during the Holocaust.

2.3.2 The Code-Mixing in Oral Histories

Pragmatic features in linguistics are another key feature described in Holocaust Testimonies, playing an essential role in understanding how survivors communicate their experiences (Table 2.1). These features explain the context and social use of language. Code-mixing, the practice of alternating between languages or dialects within speech, is a common pragmatic feature in Holocaust testimonies. The Second World War significantly impacted Europe as a whole, not just specific areas. Given the linguistic diversity of European countries, survivors used their native languages to recount their experiences, reflecting their cultural and national

Table 2.2: Code mixing Snippets extracted from Holocaust Testimonies

Languages	**Samples from testimonies**
German and Polish in English	I told Anja, 'Schnell, we must gehen to the **kryjówka—hideout**, quick!' The Gestapo was blisko, so close, and my heart was bije jak młot—beating like a hammer. *(Vladek Spiegelman, USHMM)*
German, French, and English	The SS shouted, **'Spielen, vite, play schnell!'** We tuned our violons—violins—and I whispered to Mala, **'C'est un cauchemar,** a nightmare, non? *(Fania Fénelon, USHMM)*
Yiddish, Polish, and German	They demand the kinder—children—and I must geben them. My serce, my heart, is złamane—broken. I cry, **'Oj, ratujcie**, help!' but es iz shver—too hard. *(Chaim Rumkowski, Yad Vashem)*

backgrounds. As a result, oral language in survivor testimonies consists of diverse terminology in multiple languages, reflecting the multilingual and multicultural backgrounds of Holocaust survivors. As illustrated in Table 2.3, Holocaust survivors often mix languages such as Yiddish, Polish, German, or Hebrew, each carrying unique emotional and cultural meanings. Table 2.2 highlights how the code-mixing phenomenon adds extra complexity when analysing and referencing these narratives. Moreover, this adds layers of meaning to the testimonies, making them rich sources of historical, emotional, and cultural insight. Further challenges and complexities are discussed in the section. However, the multilingual nature of Holocaust testimonies allows scholars to study language-influenced survivor storytelling, memory formation, and cultural transmission.

Speech acts are another pragmatic feature in Holocaust testimonies. Survivors use language to define events during the Holocaust, express emotions, seek validation, and establish connections with their audience in their testimonies. In testimonies, indirect speech is often used to describe harrowing and sensitive memories (Laub, 2013). For example, instead of explicitly stating, `I saw people being killed`. A survivor might say, `I saw things that no one should ever see.` This indirectness allows them to communicate the horror while managing the emotional weight of their words.

Figure 2.2: Language Distribution of Europe

Moreover, politeness strategies and culturally embedded expressions for different nationalities further shape the testimonies. Holocaust survivors use euphemisms or softened language to discuss traumatic events in their testimonies, reflecting both their emotional state and cultural norms. For example, referring to death as `passing away` or using terms like `selection` instead of `separation for execution` demonstrates how language is adapted to navigate the psychological and social impact of their experiences.

2.3.3 Emotions in Testimonies

According to a psychological perspective, language serves as a tool to manage and convey extreme feelings such as grief, fear, anger, and resilience in Holocaust testimonies. Spoken language extends beyond the meanings of words and contains non-verbal cues such as pauses and gestures that enhance emotional expressions. Table 2.4 refer to the linguistic and paralinguistic features that play a crucial part in conveying trauma and psychological distress, making oral testimonies a valuable yet complicated resource for historical and linguistic analysis (Plutchik, 1980).

Rhythm and prosody in Holocaust testimonies are other characteristics which

Table 2.3: Language Distribution across Europe

Language	**Countries**
German	Germany, Austria, Belgium, Czech Republic, Latvia
Polish	Poland, Lithuania, Ukraine, Belarus
Yiddish	Poland, France, Netherlands, Belgium, Denmark, Czech Republic, Lithuania, Latvia, Estonia, Ukraine, Belarus
Russian	Lithuania, Latvia, Ukraine, Estonia, Belarus
Dutch	Netherlands, Belgium
French	France, Belgium
Italian	Italy
Hungarian	Hungary
Rumanian	Rumania
Danish	Denmark
Czech	Czech Republic
Greek	Greece

highlight the emotional intensity of the Holocaust. Frequent pauses, silences, and self-correction in the survivor's speech underline the difficulty in verbalising traumatic experiences (Pennebaker, 2011). These disruptions are not merely linguistic irregularities but expressions of psychological states, emphasising distress, hesitation, or memory retrieval challenges. While pauses and self-corrections in speech provide psychological states of survivors, they also introduce challenges in domain-specific information extraction. In information extractions, these elements appear as "noise" in transcriptions, increasing the complexity of automatic processing and interpretation.

However, features like crying, trembling voices, sudden breaks, and vocal tremors in oral testimonies highlight the authentic emotional depth not captured in written transcripts (Felman and Laub, 1992). When analysed correctly, these raw emotional expressions provide a deep understanding of survivor trauma, revealing emotions that words alone cannot convey.

Table 2.4: Linguistic Markers of expressing the emotions

Emotions	**Linguist Marker**
Fear & Terror	Repetitions, fragmented sentences, hesitations, fast speech, code-switching
Grief & Mourning	Metaphors, long pauses, slow speech, mourning phrases
Helplessness	Passive voice, rhetorical questions, monotonous tone
Anger & Resentment	Harsh language, loud volume, direct statements, sarcasm.
Guilt	Hesitations, self-blame, fading voice
Resilience & Strength	Collective pronouns ("we, us"), strong statements, defiant speech
Gratitude & Hope	Thankful expressions, soft tone, hopeful metaphors

2.4 Computational Language Modelling

A language model (LM) is a computational model designed to understand, generate, and predict natural language. Inside the language model, the probability of a sequence of words occurring in a given order is estimated. However, the primary goal of an LM is to capture the contextual relationships between words, phrases, and sentences. Early statistical models achieved this through simple co-occurrence patterns, while more recent neural approaches based on transformer architectures have enabled deeper contextual understanding by modelling long-range dependencies across text. The purpose of LM extends across multiple NLP tasks such as text generation, text classification, named entity recognition and question & answering.

Statistical language models: aim to compute the probability of a word sequence $S = w_1, w_2, ..., w_n$, denoted as $P(S)$, by modelling the joint probability of words based on their co-occurrence patterns in a training corpus. The probability of a sequence is typically factorised using the chain rule:

$$P(w_1, w_2, ..., w_n) \prod_{i=1}^{N} P(w_i | w_1, w_2, ..., w_{i-1})$$

where $P(w_i|w_1, w_2, ..., w_{i-1})$ is the conditional probability of the $i-th$ word given its preceding context. Statistical language models employ the Markov assumption, limiting the context to a fixed number of previous words.

N-gram models are a type of statistical language model that estimates the

probability of word sequences (e.g., bigrams, trigrams) based on their frequency in a corpus. For example, in a trigram model, the probability of a word is conditioned on the previous two words.

$$P(w_i|w_1, w_2, ..., w_{i-1}) \cong P(w_i|w_{i-n+1}, ..., w_{i-1})$$

While N-gram models are simple and computationally effective, they consist of significant limitations, such as depending on a limited context window, which leads to unreliable predictions for unseen word combinations. In order to address the data sparsity issue, smoothing techniques such as Laplace and Kneser-Ney smoothing techniques were often applied with static language models. However, smoothing techniques were not able to fully resolve the underlying limitations in static language models.

In the context of Holocaust testimony analysis, statistical language models present fundamental shortcomings that make them ill-suited for the domain. The highly specialised vocabulary of Holocaust testimonies, such as comprising camp names, Yiddish and Hebrew code-mixing, and domain-specific terminologies, results in severe data sparsity when modelled using N-gram approaches. Fixed context windows further prevent these models from capturing the long-range narrative dependencies characteristic of survivor speech, where a traumatic event first mentioned early in a testimony may be contextually elaborated upon much later. Smoothing techniques offer partial mitigation, but the inherent rigidity of statistical models cannot accommodate the fragmented, emotionally disrupted syntax that is a defining feature of Holocaust oral histories. These limitations provide a direct motivation for the shift towards neural approaches better suited to the linguistic complexity of this domain.

Neural language models: represent a significant advancement by learning continuous vector representations of words known as 'word embeddings' and capturing contextual relationships through deep learning architectures. Primarily, neural language models were introduced through feed-forward neural networks,

where the goal remains to estimate the probability of a word sequence $S = w_1, w_2, ..., w_n$, denoted $P(S)$. Unlike statistical models that rely on frequency counts, neural language models model this probability as a parametrised function learned from data:

$$P(w_1, w_2, ..., w_n) \prod_{i=1}^{N} (w_i | w_1, ..., w_{i-1}; \theta)$$

Where θ represents the neural network parameters.

Neural language models capture contextual dependencies through architectures like Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks (Sherstinsky, 2020), and Transformers (Vaswani et al., 2017), which enable the modelling of longer and more complex contexts compared to statistical language models. Transformer models leverage self-attention mechanisms to process entire sequences at once, enabling them to capture long-range dependencies and nuanced semantic patterns across a text.

Compared to the statistical approaches, neural language models provide a more suitable environment for processing Holocaust testimonies, where they learn distributed vector representations of meaning, which allows word embeddings to encode semantic proximity between domain-specific terms in ways that n-gram and other frequency-based approaches cannot (Asudani, Nagwani, and P. Singh, 2023). Architectures such as LSTMs and Transformers are more capable of modelling the non-linear, emotionally fragmented structure of oral testimony (Gururangan et al., 2020). However, neural language models trained on general-domain corpora lack the historical and cultural grounding necessary for Holocaust-specific NLP tasks. This gap between general-purpose neural models and the specific demands of Holocaust testimony processing directly motivates the development and application of domain-adapted pre-trained language models, discussed in the sections that follow.

2.4.1 Encoder-Decoder Architecture Models

With the introduction of the attention mechanism by (Vaswani et al., 2017), the traditional encoder-decoder architecture has stepped into the next stage by overcoming key limitations of earlier sequence-to-sequence architectures, such as information loss in fixed-length context vectors (Bahdanau, Cho, and Bengio, 2014). Building upon this advancement, integrating self-attention within the transformer-based encoder-decoder architecture has enabled more effective modelling of long-range dependencies and contextual representations. However, applying transformer-based architectures to domain-specific, low-resource, or semantically complex tasks remains challenging because standard attention mechanisms may struggle with specific terminologies, temporal ambiguities, and implicit cultural contexts. Initially, the encoder-decoder architecture was introduced for the machine translation task. However, it has been expanded to include other NLP tasks such as text classification, spam detection, and text generation.

The above adaptability was further amplified by the advent of pretrained language models (PLMs), which leverage large-scale unsupervised learning to capture generalised linguistic patterns before fine-tuning on downstream tasks. Pre-trained models often build on the Transformer's encoder-decoder architecture, which consists of an encoder that processes input sequences to create contextual representations and a decoder that generates output sequences or labels. The self-attention mechanism in both components enables capturing long-range dependencies, making it ideal for pre-training on large, diverse datasets. Therefore, pretraining on massive corpora achieved state-of-the-art performance across different NLP tasks. Pre-trained models, such as BERT (Devlin et al., 2019), RoBERTa (Y. Liu et al., 2019), ELECTRA (Clark et al., 2020) and ModernBERT(Warner et al., 2025), employ only the encoder for tasks requiring contextual understanding, while other models, such as T5 and BART, utilise the full encoder-decoder structure to support both discriminative and generative tasks, showcasing the architecture's flexibility.

Encoder-decoder architectures have shown promising results for specific com-

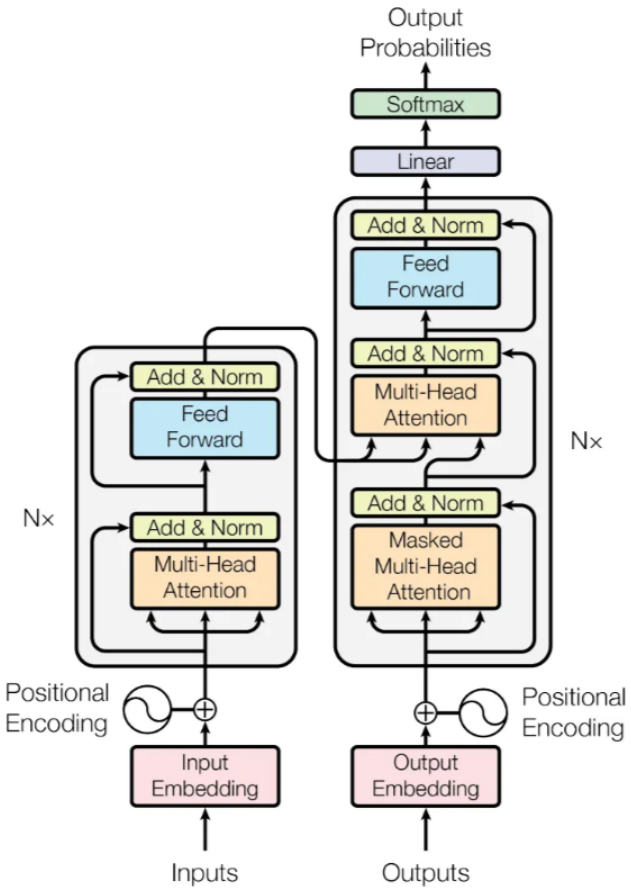

Figure 2.3: Architecture of the Transformer Model (Vaswani et al., 2017)

putational tasks related to historical data analysis (Ehrmann et al., 2023). The encoder component's capacity to build deep contextual representations of input sequences is especially valuable for NER in testimonies, where entity labels depend on rich contextual disambiguation. For example, in cases like the word *block* may refer to a residential unit, a punishment structure, or a city block depending on the surrounding context. According to this review, the encoder-only models such as BERT and RoBERTa have demonstrated strong performance on NER tasks in historical documents. However, the standard pre-training corpora underlying these models contain generic information, meaning that off-the-shelf encoder models cannot reliably recognise domain-specific terminology. This reinforces the necessity of domain-specific fine-tuning or continued pre-training on Holocaust testimony

corpora, a central methodological consideration for the present research.

2.4.2 Decoder-Only Architecture Models

With the advent of GPT models, the decoder-only models have demonstrated better performance in autoregressive text generation tasks such as language modelling, text completion, and creative writing. Decoder-only models take input, as a simple prompt or a more intricate collection of inputs, that are able to generate text. Decoder models are unique in their methodology for processing text compared to other models, which are strictly based on already existing annotated text. In that sense, decoder-only models generate "new text". The power of decoder-only models lies in their ability to not just mimic human-like text but also be creative in their responses by crafting stories, answering questions, and engaging in natural and fluid dialogue (Fu et al., 2023; C. Zhou et al., 2023).

Since decoder-only models are pre-trained on a large corpus of data encompassing a substantial portion of text available on the internet, the model has to predict the next word of each sequence of text. Once these models are pre-trained, they can be fine-tuned for a specific task. This fine-tuning can be achieved through methodologies like instruction tuning or reinforcement learning from human feedback (RHLF), which tailors the model for applications such as question-answering systems, virtual assistants, or dialogue-based systems. Regarding inferencing, decoder-only models employ algorithms such as greedy search and sampling techniques to choose the most appropriate words for generating the next part of the text. The ability to generate text makes decoder-only models effective for applications incorporating human-like interactions and content creation (D. Naik, I. Naik, and N. Naik, 2024). Several decoder-based model architectures have emerged, each offering unique enhancements:

- Autoregressive Transformers – Standard models such as GPT that predict one token based on previous tokens at a time.
- Dilated Convolutional Models – Incorporate convolution with dilation to

increase receptive fields efficiently.

- Sequence-to-Sequence GANs (Seq2Seq GANs) – Combine generative adversarial learning with sequence modelling.
- Sparse Transformers – Use sparse attention mechanisms to reduce computational cost while handling long sequences.
- Transformer-XL – Extends context length with recurrence mechanisms for better modelling of long-term dependencies.

Despite the generative strengths of decoder-only models, their application to Holocaust testimony analysis raises both opportunities and significant concerns. Their capacity to generate contextually coherent text makes them potentially suitable for tasks such as narrative summarisation, testimony completion, or question-answering over archival content (Brown et al., 2020). However, the autoregressive generation process in decoder-only models introduces a critical risk of hallucination in this domain: a model that plausibly but incorrectly generates names, dates, locations, or events within a Holocaust narrative context could cause serious harm to historical accuracy and to the integrity of survivor testimony. Furthermore, decoder-only models are pre-trained on large corpora that embed distorted, minimising, or even denialist representations of Holocaust events, introducing latent biases that are difficult to detect. These considerations suggest that while decoder-only LLMs may play a supporting role in Holocaust testimony analysis through prompt engineering approaches, their outputs require careful validation and should not be treated as historically authoritative without expert oversight.

2.4.3 Pretraining and Fine-Tuning a Language Model

Pre-training is the initial phase of learning for language models. During pre-training, models are exposed to a vast amount of unlabelled text data with the goal of capturing the underlying patterns, structures, and semantic knowledge (Radford et al., 2018; Qiu et al., 2020). This process is typically unsupervised,

allowing models to learn without explicit guidance or labelled data. A common pre-training objective is masked language modelling, where the model predicts missing or masked words within a sentence, thereby learning contextual relationships and capturing linguistic patterns (Devlin et al., 2019). Pre-trained models served as the foundation for diverse NLP tasks, leveraging Transformer-based architectures that excel in capturing long-range dependencies and contextual information (Vaswani et al., 2017).

The pre-training of encoder-only PLMs employs unsupervised objectives: Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) to learn and capture deep semantic representations from large, unlabelled corpora. In the MLM objective, a percentage of input tokens are randomly masked, and the model is instructed to predict them based on their bidirectional context, understanding of relationships and syntactic structures. The NSP objective trains the PLMs to predict whether two sentences are consecutive, enhancing their ability to maintain inter-sentence coherence. By pre-training, the above-discussed objectives enable the model to capture linguistic knowledge, such as syntax, semantics, and contextual dependencies, which serve as a foundation for transfer learning. Transfer learning in encoder-only PLMs is a key strength of such models, which allows the model to be fine-tuned with limited labelled data derived from a specific domain. During fine-tuning, the pre-trained encoder is trained for a specific domain task, such as biomedical NER (e.g., identifying gene or drug names) or legal NER (e.g., extracting contract clauses), by optimising task-specific layers added on top of the encoder's contextual embeddings. The encoder extracts domain-relevant features, such as specialised terminology. At the same time, fine-tuning adjusts the model to handle domain-specific challenges, such as limited annotated data, ambiguous entity boundaries, or complex entity relationships (nested entities in scientific texts). Models such as BioBERT (J. Lee et al., 2020) or Legal-BERT (Chalkidis et al., 2020) have demonstrated that performance improved by continuing pre-training on domain-specific corpora, aligning the model's representations with the target

domain's linguistic patterns.

Building upon the foundation of transfer learning, further pretraining is known as continued pretraining or domain-adaptive pretraining and is the process of taking a pretrained language model (such as BERT, RoBERTa, or GPT) and continuing to train it on a new dataset that is specific to a domain or task on an objective such as MLM or NSP before fine-tuning it on a specific downstream task (S. Lee et al., 2023). Further pretraining enables the model to learn domain-specific vocabulary, narrative structures, and the semantic relationships between entities. This intermediate adaptation phase is particularly crucial when the target domain differs significantly from the general corpora on which the base model was originally trained.

Fine-tuning involves adjusting the weights of a pre-trained model on a domain-specific dataset to optimise its performance for a particular task, such as NER (Anisuzzaman et al., 2025). During fine-tuning for NER, the model learns to identify and classify entity boundaries and types within the texts, leveraging both the general linguistic knowledge from pretraining and the domain-specific understanding acquired during further pretraining. Fine-tuned models improve the performance of the LLM on the specific task or domain by adjusting the weights of the model to better fit the data. Supervised fine-tuning (SFT) uses labelled data to train the LLM, which contains pairs of input and output data for a specific domain or task. Reinforcement Learning from Human Feedback (RLHF) incorporates human evaluators' judgements as an alignment fine-tuning approach which acts as an optimisation process, guiding the model toward preferred outputs (Ouyang et al., 2022). Although RLHF can yield higher-quality results than SFT, it is often considered more resource-intensive and operationally complex due to the requirement for extensive human annotation and iterative reward modelling (Mei et al., 2025)

.

Beyond these approaches, specific techniques such as hyperparameter tuning, parameter-efficient fine-tuning and in-context learning are used for fine-tuning LLMs

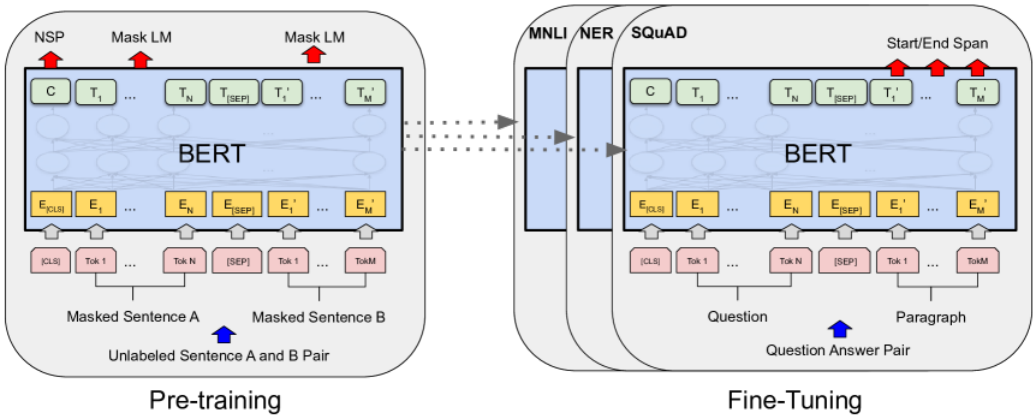

Figure 2.4: Pre-Training and fine-tuning processes of Large Language Models (Devlin et al., 2019)

for specialised use cases (Anisuzzaman et al., 2025).

- In the in-context learning approach, the model's ability to learn and perform a specific task is based on a few examples or instructions embedded directly within the input context, eliminating the need for explicit fine-tuning or parameter updates. As a result, model output is generated purely through contextual cues in the prompt, operating entirely without gradient-based training or modifications to the model's weights (H. Liu et al., 2022).

- Hyperparameter tuning involves systematically adjusting basic training parameters such as learning rate, batch size, optimiser, and number of epochs to optimise a model's performance. Hyperparameters define the model's learning behaviour, influencing key aspects such as convergence speed, training stability, and stopping criteria. By iteratively refining these settings, it's able to balance training efficiency with model accuracy, ensuring the model learns effectively without overfitting or underfitting.

- Parameter-efficient fine-tuning (PEFT) is an efficient technique that selectively updates only a small subset of model parameters during fine-tuning, introducing task-specific layers or modifying existing ones. PEFT significantly

reduces computational and storage costs, keeping performance comparable with complete fine-tuning. Some PEFT techniques are low-rank adaptation (LoRA) (E. J. Hu et al., 2022), quantised low-rank adaptation (QLoRA) (Dettmers et al., 2023), prefix tuning, and prompt tuning. LoRA is able to update the model weights and compress a version of the LLM weights (E. J. Hu et al., 2022). While the compression may lose some data, under the assumption that many of the model weights are redundant, it leads only to a small decrease in performance relative to savings in memory and required compute power.

With the field of digital humanities growing, the development of BERT-like models trained on 18th-century and 19th-century historical data has accelerated. In an early pioneering effort, Hosseini et al. (2021) and Beelen et al. (2021) took a significant step by further training a standard BERT model, utilising English books published between 1760 and 1900, along with data from the Oxford English Dictionaries. The BERT model was pretrained from scratch, using the Eighteenth Century Collections Online (ECCO) dataset, which encompasses a vast collection of over 180,000 titles published during the eighteenth century. Another model, MacBERTh Manjavacas and Fonteyn (2021) was also pre-trained from scratch and evaluated for various tasks such as Part-of-Speech (POS) tagging, Named Entity Recognition and Word Sense Disambiguation. The authors of this study further explained that the pre-training approach exceeds the capabilities of the standard BERT. In addition to historical documents available in English, a multilingual BERT model has been developed employing French historical documents, following the same approach as MacBERTh. Moreover, another BERT-based model (BERToldo) was developed using Italian books and Wikisource, which was evaluated using Dante Alighieri's works for POS tagging tasks (Palmero Aprosio, Menini, and Tonelli, 2022).

For Holocaust testimony analysis, the pretraining and fine-tuning paradigm offers a solid approach for overcoming the domain adaptation problem. General-

purpose pre-trained models lack exposure to the specific linguistic features, historical vocabulary, and multilingual characteristics of Holocaust testimonies (Xiaochuang Han and Eisenstein, 2019). But having continued pre-training on testimony-specific corpora is able to infuse domain-relevant representations before fine-tuning on labelled tasks such as NER or relationship extraction (Villena, Bravo-Marquez, and Dunstan, 2025).

2.4.3.1 Instruction Learning

Instruction learning is a methodology that improves the performance and generalises the capabilities of LLMs through explicit instructions (Chung et al., 2022). Rather than being trained on task-specific datasets with fixed formats, instruction learning allows for describing different tasks through natural language prompts or instructions to the model. The diversity of tasks and instruction formats teaches the model to generalise across domains and adapt to novel tasks described in natural language. Instruction learning serves as a foundation for prompt engineering, which involves designing input prompts to retrieve the desired outputs from the LLM without modifying its original weights. For example, studies have shown that structuring prompts into components such as Role (e.g., "You are a historian"), Rule (e.g., "Explain in simple terms"), and Task (e.g., "Describe the causes of World War I") leverages the model's instruction-following ability to produce contextually aligned responses (H. Li et al., 2025). To achieve generalisation, instruction learning is implemented through a specific fine-tuning process known as instruction tuning, where a pretrained LLM is trained on curated datasets of instruction-output pairs (J. Wei, Bosma, et al., 2022). In contrast to the traditional fine-tuning mechanism, instruction tuning uses different datasets or models encompassing tasks such as question answering, summarisation, code generation, and reasoning. This diversity enables the model to learn, interpret and respond to varied instructions, enhancing its ability to handle novel prompts effectively. The following section thoroughly explores prompt engineering, focusing on techniques such as zero-shot, few-shot,

and chain-of-thought prompting.

In the Holocaust testimony domain, instruction learning provides a method for guiding LLMs toward historically and ethically sensitive tasks without the need for large quantities of labelled data. Structuring prompts with role-based instructions combined with explicit task rules has been shown to produce more contextually grounded outputs than normal task prompts. Instruction tuning on diverse tasks also enhances the model's capacity to handle the structural heterogeneity of oral testimonies, which blend narrative, descriptive, and evaluative discourse modes in ways that resist simple classification schemas. Nevertheless, instruction-tuned models must be evaluated carefully in this domain to ensure that task performance does not come at the cost of misrepresenting or oversimplifying the complexity of survivor accounts.

2.4.3.2 Prompt Engineering

Prompt engineering is a technique used in LLMs to optimise the performance by designing effective input prompts to retrieve outputs without modifying the model's weights. Prompts are natural language instructions or queries that humans use to interact with LLMs, guiding their responses to achieve specific outcomes (Sander Schulhoff et al., 2024; Reynolds and McDonell, 2021). By employing prompt engineering strategies such as zero-shot, few-shot, and chain-of-thought prompting, users are able to utilise the generalisation capabilities of LLMs with improved accuracy and alignment to user intent (Sahoo et al., 2024; J. Wei, Xuezhi Wang, et al., 2023).

- Zero-shot learning: It involves providing an LLM with a task description in natural language without including any example inputs or outputs(J. Wei, Bosma, et al., 2022).Moreover, zero-shot prompting leverages the model's ability to perform in-context learning. The model uses pretrained knowledge and fine-tuned instructions for particular tasks while mapping the input prompt to an appropriate output.

- Few-shot learning: It relies on the model's ability to perform in-context learning, where the prompt acts as a mini-training dataset by including a small number of example input-output pairs in the prompt to clarify the expected response format, style, or content at inference time (Brown et al., 2020). Instruction tuning enables the model to adapt to new examples for diverse task formats without further weight updates. The model's attention mechanism prioritises relevant patterns in the examples to inform its response.

- Chain-of-Thought Prompting: CoT prompting utilises the model's ability to generate coherent sequences, enhanced by instruction tuning on tasks requiring sequential reasoning, such as logical, mathematical, or multi-step reasoning (J. Wei, Xuezhi Wang, et al., 2023). During inference, the model uses its autoregressive nature to build on each reasoning step, with the attention mechanism focusing on relevant parts of the prompt to maintain logical consistency.

Beyond the above-discussed techniques, other advanced strategies such as role specification ("Act as a historian") or output constraints ("List three key points in bullet format") further refine model outputs. Prompt engineering bridges the gap between raw model capabilities and practical applications, offering a scalable way to customise LLM behaviour for different domains. However, domain-specific tuning is required to further understand model limitations, such as sensitivity to prompt phrasing or the risk of hallucinated responses.

Prompt engineering is a particularly tractable approach for Holocaust testimony analysis given the scarcity of large, task-specific annotated datasets in this domain. Zero-shot prompting, which relies solely on the model's pre-trained knowledge, is limited by the under-representation of Holocaust-specific content in standard training corpora; a model prompted to identify concentration camp names or perpetrator organisations may perform inconsistently when encountering less prominent sites or non-German language variants. Few-shot prompting addresses this partially by demonstrating to the model the expected output format with

carefully selected Holocaust-specific examples, improving both entity recognition and relation extraction in this context. Chain-of-thought prompting may be particularly valuable for tasks requiring implicit reasoning over testimony content, such as inferring a survivor's likely location based on fragmentary geographical references or determining temporal sequences from non-linear narrative accounts. However, sensitivity to prompts is a known limitation of all prompt engineering approaches, and in the context of testimony analysis, slight differences in wording could produce historically misleading outputs. Therefore, systematic evaluation across prompt variants is essential.

2.4.3.3 Retrieval Augmented Generation (RAG)

Decoder-only models, or pretrained LLMs known as Foundation Models (FMs), have demonstrated impressive success in language generation applications because of their ability to produce fluent and contextually relevant text (C. Zhou et al., 2023). However, these models' parameters are static and store temporal knowledge. Once trained on massive amounts of text data, the models cannot access new or external information without being retrained. As a temporal limitation, pretrained LLMs were less effective for tasks that require the latest information or domain-specific information, leading to factual inaccuracies (hallucinations) and biases (L. Huang et al., 2025).

To overcome the above-discussed limitation, the Retrieval-Augmented Generation (RAG) approach was introduced (P. Lewis, Perez, Piktus, Petroni, Karpukhin, Goyal, Küttler, M. Lewis, Yih, Rocktäschel, et al., 2020b). RAG combines retrieval-based (a retriever) and generative models (a generator) to provide additional information from an external knowledge source separated from the LLM's reasoning capability, which can be easily accessed and updated. The retriever component searches a large corpus of documents to find relevant information based on the input query, while the generator component, often a sequence-to-sequence model such as GPT or BERT, takes the retrieved documents and the original query to generate a

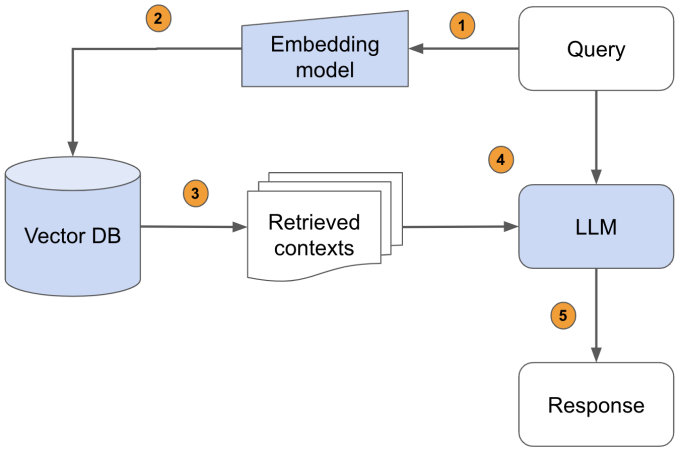

Figure 2.5: Architecture of the Retrieval-Augmented Generation (RAG) Approach

coherent and contextually accurate response.

The core components of the RAG approach consist of:

- Parametric knowledge: Learned during training that is implicitly stored in the neural network's weights.
- Non-parametric knowledge: Stored in an external knowledge source, such as a vector database.

Parametric knowledge is intrinsic to the model and is shaped during the extensive training process with large datasets. Moreover, parametric memory is notable for its ability to grasp linguistic nuances and complex contexts, but it has limitations in terms of knowledge updating and the explainability of its decisions. Non-parametric knowledge is mainly employed in information retrieval systems, and it is easily updatable and transparent, as the knowledge used in generating answers can be directly inspected.

The following Figure 2.5 describes the RAG approach.

- Pass the query to the embedding model to semantically represent it as an embedded query vector
- Pass the embedded query vector to the vector DB.
- Retrieve the top-k relevant contexts, measured by distance between the query embedding and all the embedded chunks in our knowledge base.

- Pass the query text and retrieved context text to our LLM.
- The LLM will generate a response using the provided content.

The RAG performs effectively in low-resource settings and highly specialised domains, where access to relevant contextual information is crucial (Choi et al., 2025). RAG offers a practical and scalable solution for knowledge-intensive NLP applications by enabling models to look up the information in real time. As explained in section two, the RAG approach consists of three main components.

1. **Retrieve:** The query retrieves relevant context from an external knowledge source. In the process, the query is embedded with an embedding model into the same vector space as the additional context in the vector database to perform a similarity search, and the top k closest data objects from the vector database are returned.
2. **Augment** The query and the retrieved additional context are stuffed into a prompt template.
3. **Generate** Finally, the retrieval-augmented prompt is fed to the LLM.

However according to the (Y. Gao et al., 2023) normal approach of RAG or the naive RAG approach presented some drawbacks such as

- The retrieval phase often struggles with precision and recall, leading to the selection of misaligned or irrelevant chunks and the missing of crucial information.
- Generation models might overly rely on augmented information, leading to outputs that echo retrieved content without adding insightful or synthesised information.

To overcome the above limitations, an advanced RAG was introduced. In the Advanced RAG approach design, limitations were handled in different stages: 1) pre-retrieval, 2) retrieval, and 3) post-retrieval optimisations.

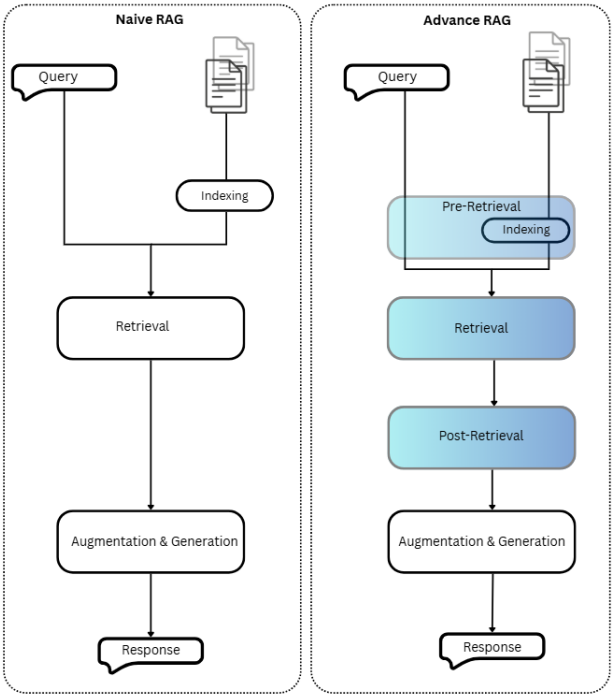

Figure 2.6: Encoder-Only Transformer Architecture for Holocaust Named Entity Recognition

1. Pre-retrieval stage: This stage focuses on optimisation through data indexing and query enhancements, which aim to structure data in a way that maximises retrieval efficiency. Techniques such as sliding window integration between data chunks and the addition of metadata (e.g., timestamps, purpose tags, or chapter identifiers) significantly improve the pre-retrieval phase. Furthermore, pre-retrieval optimisation is not limited to indexing strategies; it could be employed with inference-time techniques, including query routing, query rewriting, and query expansion.

2. Retrieval optimisation stage: The retrieval stage identifies the most relevant context based on vector search, which calculates the semantic similarity between the query and the indexed data. Thus, the majority of retrieval optimisation techniques are based on optimising the embedding models to improve the accuracy of the RAG approach.

Fine-tuning embedding models: This customises embedding models to domain-specific contexts with novel or rare terms.

In order to improve the accuracy of a decoder-only model, fine-tuning the underlying embedding model can be an effective solution. This process allows the model to adapt to domain-specific language and patterns, enhancing its ability to generate relevant and context-aware responses. However, full fine-tuning of large embedding models (such as those used in transformer architectures) is computationally expensive and time-consuming, requiring substantial GPU resources, large annotated datasets, and careful parameter optimisation. Additionally, it leads to overfitting if not done carefully, especially in low-resource domains. To mitigate these issues, recent approaches like Parameter-Efficient Fine-Tuning (PEFT), such as LoRA (Low-Rank Adaptation) or adapter layers, are increasingly adopted, allowing targeted adaptation with significantly fewer trainable parameters and reduced computational cost.

Dynamic Embeddings: This adapts to the context in which words are used, unlike static embedding, which uses a single vector for each word. Besides vector search, combining vector search with keyword-based search (hybrid search) is often used in the retrieval optimisation stage.

3. Post-retrieval optimisation stage: At this stage, the retrieved context may present challenges, such as exceeding the context window limit or introducing noise, which can hinder the critical information. To mitigate these issues, post-retrieval optimisation is performed using two key techniques: prompt compression and re-ranking.

The Retrieval-Augmented Generation framework performs well with the Holocaust testimony analysis, where the combination of domain-specific information and contextual language generation is critical. A RAG approach built over a curated knowledge base of Holocaust archival data, incorporating testimony transcripts, geographical databases, and historical encyclopaedias, allows an LLM

to ground its responses in verified historical content rather than relying solely on potentially incomplete or biased parametric knowledge (P. Lewis, Perez, Piktus, Petroni, Karpukhin, Goyal, Küttler, M. Lewis, Yih, Rocktäschel, et al., 2020b). Moreover, the non-parametric knowledge stored in a RAG system can be updated as new archival material is digitised, making it a scalable solution for an ongoing preservation effort. The advanced RAG framework's pre-retrieval query optimisation techniques are also important, given that testimony queries are often implicit, emotionally mediated, and linguistically fragmented (Y. Gao et al., 2023). A further evolution beyond Advanced RAG is the integration of knowledge graphs as the underlying index structure. (Edge et al., 2024) propose GraphRAG, a graph-based approach to query-focused summarisation over large, private text corpora that addresses a fundamental limitation of vector-based RAG: its inability to support global sensemaking — that is, answering broad, thematic questions that require synthesising information distributed across an entire document collection rather than retrieving a handful of locally relevant chunks. GraphRAG operates in two pre-indexing stages: first, an LLM extracts an entity knowledge graph from the source documents, where nodes represent key entities and edges encode relationships between them; second, graph community detection algorithms (e.g. the Leiden algorithm) partition the graph into clusters of closely related entities, for which the LLM pre-generates hierarchical community summaries. At query time, each relevant community summary is used to generate a partial response, and all partial responses are aggregated into a final answer. In evaluations on corpora of approximately one million tokens, GraphRAG yielded substantial improvements over a conventional RAG baseline in both the comprehensiveness and diversity of generated answers (Edge et al., 2024). This approach is particularly relevant to Holocaust testimony analysis, where queries frequently concern broad thematic patterns such as the geographical spread of deportations, recurring perpetrator networks, and systemic descriptions of camp conditions that span hundreds of testimonies simultaneously and cannot be adequately addressed by retrieving a small number of individually

similar passages.

2.4.4 Computational Applications for Oral Narratives

In recent years, Holocaust testimony transcripts have received relatively little attention, and a limited number of computational techniques have been developed to analyse and extract structured information from unstructured information. Because of that, the following section will mainly focus on the information extraction techniques and their usage for historical documents, including Holocaust testimonies.

Information Extraction (IE) is a fundamental process in Natural Language Processing (NLP) that automatically identifies, extracts, and structures useful information from unstructured or semi-structured text (S. Singh, 2018). This extracted information can then be used for knowledge discovery, data analytics, and decision-making across various domains, including digital humanities (Atanassova, Bertin, and Mayr, 2022).

2.4.5 Evolution of Named Entity Recognition

Named entity recognition (NER) is a widely used information extraction technique across different domains (Pakhale, 2023). Previous research has leveraged NER systems to extract domain-specific factors using machine learning algorithms (Kumar and Starly, 2022; Nenno, 2024). With the evolution of technology, NER has emerged as a core NLP task, adapting new technology by improving accuracy. NER has been applied to various forms of social media text, including YouTube transcriptions and podcasts. In the earlier stages of development, these tasks primarily relied on Conditional Random Fields (CRF) and rule-based models for entity extraction (Hatmi et al., 2013; Pearson, Seliya, and Dave, 2021). With the technological evolution, researchers have adopted machine learning techniques to improve the extraction of diverse types of Named Entities (NEs). This evolution has led to the development of models capable of recognising not only general named entities but also domain-specific entities, tailored to the unique linguistic

and contextual characteristics of specialised fields such as biomedical records, legal documents, and historical narratives (Zaghloul and Trimi, 2017).

With the development of deep learning algorithms over the last decade, NER systems have been adapted to deep learning mechanisms (Yadav and Bethard, 2018). As a result, they have obtained better performance on both structured and unstructured texts by effectively understanding contextual meaning. Extraction of named entities from unstructured text is a challenging task because of the complexity of the language, variations of the context and the inconsistency of the redefined structure of the text (Zaghloul and Trimi, 2017). However, deep learning algorithms' capability of understanding contextual meanings has improved performance in identifying complex domain-specific NE tags from unstructured texts (Yadav and Bethard, 2018).

Domain-specific NE tags depend on the particular task and the type of information you aim to extract from the text. General NER systems were developed to extract tags such as Person (PER), Geopolitical Entity (GPE), Location (LOC), Organisation (ORG), Language (LAN), etc. However, for domain-specific tasks that require extracting more content-specific information beyond the capabilities of the general NER system, fine-tuning with a custom annotated dataset is necessary for the specific task. As a result, domain-specific NER models have emerged and developed for information retrieval across diverse domains. Medical NER systems (Yonghui Wu et al., 2018; Dunstan et al., 2024; Pearson, Seliya, and Dave, 2021) have been developed to identify entities such as diseases, medications, and symptoms in healthcare and biomedical research, legal NER (Leitner, Rehm, and Moreno-Schneider, 2019) systems have been designed for extracting case laws, legal statutes, and contractual entities and financial NER (Yuzhe Zhang and H. Zhang, 2023) systems for identifying company names, monetary values, and stock-related terms are now widespread and used for knowledge discovery in their respective fields.

In the domain of historical text and document processing, previous research has been conducted on archives (Poso et al., 2023; Cunha and Ramalho, 2021), historical

maps (Karsvall and Borin, 2018), historical books (novels, poems) (W. Jiang, 2024), directories (Abadie et al., 2022), newspapers (Ruokolainen and Kettunen, 2018; Neudecker, 2016) and oral testimonies. However, we have narrowed down our scope only to explore unstructured textual documents. Therefore, we focus on the NER approaches applied to unstructured historical texts such as historical newspapers and oral testimonies.

2.4.5.1 Tools and Technologies Used for NER in Historical Documents

According to the computational approaches experimented with in previous studies of NER, techniques can be classified into rule-based heuristics and machine-learning approaches (Keraghel, Morbieu, and Nadif, 2024; Pakhale, 2023).

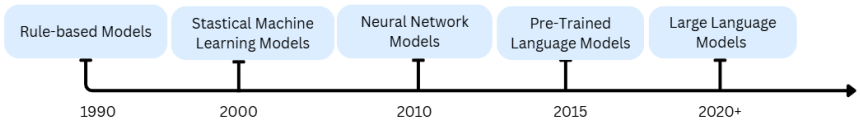

Figure 2.7: Timeline of Named Entity Recognition Evaluation Methods

In early days, domain-specific NER systems mostly relied on rule-based techniques. Most of these frameworks were based on handcrafted linguistic rules and regular expressions tailored to a particular terminology related to the specific domain. spaCy is one of the NER tools used for extracting domain-specific named entities (Honnibal and Montani, 2017). Even though the spaCy NER system is a rule-based model designed for extracting general named entities, adding specific rules and retraining the model enables it to extract domain-specific tags from the texts. spaCy has various pretrained models that can be used for NER. Earlier versions were based on the traditional CNN-based statistical model, which used statistical word vectors like GloVe on general English language text corpora. However, after introducing the transformer architecture, spaCy released a transformer-based pretrained model based on contextual word embeddings on massive text corpora such as Wikipedia and web text. However, fine-tuning with domain-specific rules is

required for domain-specific applications. Stanza (Qi et al., 2020)is another NER tool that the Stanford NLP Group introduced for natural language processing in 2020, which was developed using neural network architecture on top of the PyTorch framework. Stanza works with 70+ languages for different NLP tasks, including NER, by providing domain-specific fine-tuning. However, in contrast to domains such as medical, legal, and financial, the development of NER frameworks for analysing historical documents has received comparatively less attention because of the scarcity of annotated historical datasets, variations in context-specific language, and the complexities of document preservation and digitisation (Ruokolainen and Kettunen, 2018; Neudecker, 2016; Ehrmann et al., 2023; De Toni et al., 2022).

Deep learning approaches to NER, particularly through fine-tuned transformer-based language models, have emerged as state-of-the-art techniques in recent years. As a result, most of the domain-specific NER systems have employed encoder-only architectures such as BERT (Devlin et al., 2019), RoBERTa (Y. Liu et al., 2019), ELECTRA (Clark et al., 2020), and T5 (Raffel et al., 2020) due to their high performance and have been extensively trained on specific sources (Ehrmann et al., 2023). Moreover, through fine-tuning, these models demonstrate enhanced performance in recognising named entities (Aguilar, 2022; Blouin et al., 2021; Gruber et al., 2024).

The emergence of decoder-only transformer architectures, which are called Large Language Models (LLMs), has enhanced the performance and accuracy of various NLP tasks, including the NER, (Boyu Liu, 2002). These models predict tokens sequentially, allowing for a deeper contextual understanding of language and large-scale pretraining on diverse datasets to recognise entities that belong to different domains. Moreover, the large language models such as GPT-4 (Radford et al., 2018), LLaMA (Touvron et al., 2023) and Mistral (A. Q. Jiang et al., 2023) models were fine-tuned using various prompt engineering techniques such as few-shot or zero-shot learning to reduce the requirement of labelled domain-specific datasets (Pornprasit and Tantithamthavorn, 2024; Shin et al., 2023). Moreover, different studies have

examined historical newspapers in multilingual contexts, highlighting the linguistic diversity and LLMs' applicability for NER(Toni et al., 2022; González-Gallardo et al., 2023; Santos et al., 2024; Sarker et al., 2024). Additionally, (González-Gallardo et al., 2023) and (Hiltmann et al., 2025) have integrated different LLMs with prompt engineering techniques to recognise and classify named entities in the historical documents.

A significant recent development bridging fine-tuned encoder models and full LLMs is the emergence of generalist, lightweight models for zero-shot NER. (Zaratiana, Tomeh, et al., 2024) propose GLiNER, a compact bidirectional transformer encoder that recognises arbitrary entity types at inference time without task-specific fine-tuning. By encoding entity labels and input text in parallel rather than generating tokens sequentially, GLiNER runs efficiently on standard CPU hardware and matches or surpasses both ChatGPT and fine-tuned LLMs in zero-shot benchmarks despite its smaller size — making it well suited to low-resource, domain-specific settings such as Holocaust testimony analysis, where annotated data is scarce and general-purpose NER systems do not cover entities such as camp names, perpetrator organisations, or deportation routes. (Zaratiana, Pasternak, et al., 2025) extend this work with GLiNER2, a unified schema-driven framework that adds text classification and hierarchical structured data extraction to the original NER capability within a single efficient model, directly addressing the common practice of deploying separate models for each extraction task in testimony analysis pipelines. However, some studies reported that due to overgeneralisation and hallucination issues, LLMs performed worse in many or even all tasks they were given, especially compared to BERT models (González-Gallardo et al., 2023; Sarker et al., 2024). While, other studies explained that further experiments are required for further understanding of the full capability of LLMs for the NER tasks (Toni et al., 2022; Santos et al., 2024; Hiltmann et al., 2025).

2.4.6 Evolution of Relationship Extraction

Relationship Extraction (RE) from the text is considered one of the complex tasks in NLP (X. Zhao et al., 2024). Studies have proven recognising semantic relationships between entities in structured texts is easier than in unstructured texts (Hsu et al., 2015). However, studies related to RE have been thriving in recent years (Cui et al., 2017; Xin Xu et al., 2022). Surveys have reported that most studies of RE were based on the before DL techniques emerged and focused primarily on rule-based or statistical approaches (Pawar, Palshikar, and Bhattacharyya, 2017; Zelenko, Aone, and Richardella, 2003). (Cui et al., 2017) focused on both traditional RE approaches and DL-based approaches. However, given the time frame, recent DL approaches were omitted from that study. Moreover, (Xin Xu et al., 2022) focused on the low-resource RE problem, while (Bassignana and Plank, 2022) discussed RE datasets and scientific relation classification approaches. As a recent study (X. Zhao et al., 2024) has explored DL-based approaches for RE and introduced a new taxonomy to categorise existing works from three perspectives (text representation, context encoding, and triplet prediction).

In theory, RE involves a pipeline of steps as illustrated in Figure 2.8. From the raw text, the pipeline identifies the entities and eventually assigns the NER/Mention Detection (MD). After entities are identified, different approaches were followed to extract RE from different angles.

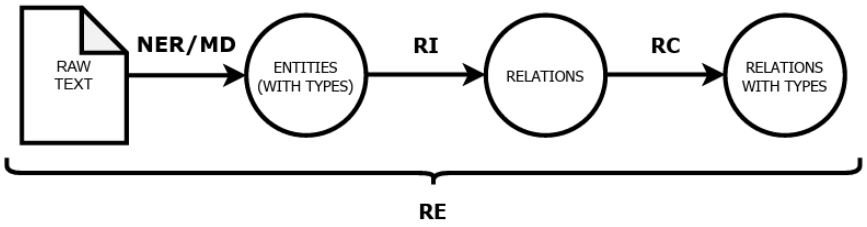

Figure 2.8: Relationship Extraction Pipeline (Bassignana and Plank, 2022)

RE involves two primary subtasks: Relation Identification (RI), which recognises the relationships between entity pairs via a binary classification task, and Relation

Classification (RC), which categorises specific types of relationships into predefined classes (Ye et al., 2019). One standard methodology for RE is triplet prediction, which jointly addresses subtasks by detecting entity boundaries and classifying the relation types while forming structured (head, relation, tail) triples from unstructured text.

Pipeline-based Approach: In the pipeline-based approach, separate the extraction of entities and relations (Miwa and Bansal, 2016; Son et al., 2022). In the first stage, all candidate entities in the sentence are annotated manually or identified through the NER models, and a classifier is used to determine the relation between every possible pair of identified entities.

The Span-Based Approaches: The span-based approaches process each sentence into spans and perform span classification to obtain predicted entities (Dixit and Al-Onaizan, 2019). Previous studies utilised pre-trained Transformers (BERT) to convert words into embeddings, compute span and relation representations, handle classification tasks, and create contextual meanings using different attention methods (Eberts and Ulges, 2020; B. Ji et al., 2020).

The Seq2Seq-Based Approaches: The Seq2Seq-based approaches receive unstructured text as input and directly generate head, relationship and tail (Nayak and Ng, 2020; R. H. Zhang et al., 2020). The Seq2Seq-based approach is able to extract relations of any length by treating them as text generation tasks. However, complexities were reported in the process of entity boundary detection.

The MRC-Based Approaches: Machine reading comprehension (MRC)-based approaches treat the entity RE task as a multi-turn QA task (X. Li et al., 2019; T. Zhao et al., 2021). (Levy et al., 2017) has formulated the RE task as a QA task, where the relations are defined by natural-language question templates. Further, (X. Li et al., 2019) and (T. Zhao et al., 2021) transformed the RE task into a multi-turn QA task, providing a natural way to identify the entities and relations in a sentence. The process then involves analysing text passages and answering content-specific questions.

2.4.6.1 Tools and Technologies Used for Relationship Extraction

Recent advances in RE have largely employed deep learning techniques. The technical methods employed for identifying semantic relationships between entities were classified into two methods: Text representation and context encoding.

Text representation converts text into a real-valued vector. In the vector space, similar meanings are expected to be closer in the vector space. Primarily, text representation learning approaches related to RE were classified into word-level (S. Zheng et al., 2017; P. Zhou et al., 2017), character-level, position-level (Zeng et al., 2014), and syntactic-level (Kun Xu et al., 2015) embedding representations. Individual embedding represents specific aspects of textual information, while hybrid embeddings combine multiple properties of embeddings to capture a wider range of linguistic features, thereby enhancing the overall representation quality for RE tasks. Additionally, the selection of external features depends on the specific application requirements.

Context encoding focuses on extracting lexical-level features from the information. Moreover, able to learn sentence-level features by capturing contextualised meaning. Prior to the pre-trained models emerging, context encoding was implemented with the support of neural network architectures, such as CNNs (Yatian Shen and X.-J. Huang, 2016), RNNs and attention-based neural networks.

Recently, PLMs (Qu et al., 2020; T. Zhao et al., 2021) have achieved promising results in modelling RE problems. The majority of PLMs were trained on large-scale corpora and fine-tuned on annotated task-specific data to adapt semantic information for RE (Soares et al., 2019; Du et al., 2018). Particularly in scenarios with limited data availability, fine-tuning PLMs on the target task has proven to be an effective practice. However, being solely based on PLMs often causes challenges when identifying new and domain-specific relations due to extensive data annotation, which can be time-consuming and labour-intensive (Gharagozlou et al., 2023).

Prompt tuning (X. Chen et al., 2022; Son et al., 2022) introduced a new paradigm to RE by bridging the gap between pre-training and fine-tuning downstream RE

tasks. Recent works (Xu Han et al., 2022; Schick, Schmid, and Schütze, 2020) show that prompt learning has leveraged the capabilities of the PLM to perform a specific task by adapting it to the target task through training on a smaller, task-specific dataset. The main drawbacks of PLMs are their resource-intensive nature, requiring significant computational power for both training and inference, and their tendency to overfit on smaller or domain-specific datasets.

LLMs showcase impressive generation capabilities and offer promising new directions to overcome the bottleneck of manually labelling documents, particularly in automatically identifying new relationships. Even with the limited annotations, LLMs enhance performance in relation extraction tasks, utilising their memorisation of vast textual patterns and their emergent reasoning capabilities. Additionally, few-shot and zero-shot prompt learning allow RE systems to extract relationships with unlabelled or a few labelled data (Wittgen, Hasibi, and Thill, 2023). However, the inference latency and financial cost associated with calling LLMs' APIs are higher than fine-tuning PLMs.

While general RE models were trained on open-domain corpora, domain-specific RE was developed to identify semantic relationships within the domains such as medical (Q. Wei et al., 2020), financial and legal (W. Xu et al., 2022). For domain-specific RE, traditional RE methods, such as Subject-Object-Verb (SOV) extraction and argument extraction, have been used, particularly in structured and semi-structured texts like Wikipedia articles (W. K. Lee et al., 2012; Jiahui Wang, Yue, and Duan, 2023). Most methods were based on syntactic parsing, dependency trees, and rule-based approaches used in linguistics. However, these methods perform poorly in unstructured and noisy text, such as spoken language transcripts and social media conversations, where sentence structures are highly variable (Subramaniam et al., 2009). With the emergence of deep learning and PLMs, domain-specific RE has adapted to identify complex relationships in structured and unstructured texts (T. Zhao et al., 2021). Moreover, deep learning techniques have combined with the rule-based approach, being a hybrid approach to improve the accuracy of RE (K. Wu

et al., 2023). Additionally, learning contextual meaning in pre-trained language models enables the capture of complex semantic relationships between entities across the multilingual data in diverse domains (Papanikolaou, I. Roberts, and Pierleoni, 2019). Mirroring the development of GLiNER for zero-shot NER, (Boylan, Hokamp, and Ghalandari, 2025) introduce GLiREL (Generalist Lightweight model for zero-shot Relation Extraction), extending the same bidirectional transformer paradigm to the RE task. Rather than treating each entity pair and candidate relation label as a separate input, an approach that becomes computationally prohibitive when many entities and label types are present, GLiREL encodes all relation labels and entity pairs simultaneously in a single forward pass, computing similarity scores between every label embedding and every entity pair representation at once (Boylan, Hokamp, and Ghalandari, 2025). GLiREL achieves state-of-the-art zero-shot relation classification performance. However, adapting RE for historical contexts has been rarely explored in previous research using deep learning mechanisms (S. Yang et al., 2023; Haris, Cohn, and Stell, 2024), leaving significant opportunities for uncovering nuanced connections within historical texts.

2.4.7 Knowledge Modelling and Visualisations

Tracing the paths of the Holocaust is one of the critical concerns related to a significant historical event. The movement of survivors and victims was subjected to forced deportations, escapes, or resettlement during this period and considered as vital information when reconstructing historical narratives, uncovering personal stories, and preserving collective memory (Hirsch and Spitzer, 2009). According to most narratives, individuals were located across multiple geographical locations, such as concentration camps and ghettos, hiding places, refugee camps, and post-war resettlements. Mapping the dots of this journey is important for historical research because it provides in-depth knowledge about survival strategies, the logistics of persecution, and the impact of displacement on different communities (Knowles, Cole, and Giordano, 2014).

Knowledge bases utilise the graph-based representations, offering a powerful framework for structuring and querying complex data. In such KBs, entities such as names, locations, or events are represented as nodes, while relations between entities form edges, creating a knowledge graph that captures the interconnected nature of Holocaust narratives. These graphs, often encoded using standards such as RDF (Resource Description Framework), enable semantic queries to trace paths across time and space, revealing patterns in deportations or survival networks (Hogan et al., 2021). By integrating automated information extraction techniques, such as NER and RE, KBs will populate from unstructured texts, enhancing scalability and depth in historical reconstruction (B. Chen and Bertozzi, 2023). This approach not only supports historians in mapping individual journeys but also preserves the semantic richness of Holocaust data for future generations.

As another method of visualisation, spatial analysis of Holocaust survivor testimonies has the ability to map fragmented data sources (Ezeani, Rayson, I. N. Gregory, et al., 2024). Many oral testimonies describe uncertain or imprecise locations where survivors have been. However, when cross-referenced with historical maps, transport records, and administrative documents, Geographic Information Systems (GIS) can help establish probable movement trajectories (Wagner, Keydar, and Abend, 2023). With the advancement of technology, GIS systems have adopted deep learning and LLMs in the context of historical data (Ståhl and Weimann, 2022). Compared to other areas of NLP, previous research on Holocaust-related content has primarily focused on integration with GIS systems (Schierman, 2025).

2.4.7.1 Knowledge Representation

The capabilities of LLMs in complex NLP tasks, such as reasoning and relationship extraction, enable the automated construction of knowledge graphs from unstructured text. These state-of-the-art models are able to identify entities, interrelationships, and structure information efficiently with less human effort (Y. Zhu et al., 2024). Moreover, this advancement allows knowledge graphs to be

dynamically updated and expanded as new information becomes available, making them more practical and scalable for real-world applications. A knowledge graph is fundamentally composed of three core components: entities, the relationships between them, and communities (or clusters) of interconnected nodes.

- An entity is a distinct object, person, place, event, or concept that has been extracted from a chunk of text. Entities form the nodes of the knowledge graph. During the formulation of the knowledge graph, duplicates are merged while preserving their various descriptions, creating a comprehensive representation of each unique entity.

- A relationship defines a connection between two entities in the knowledge graph. These connections are extracted directly from text, alongside entities. Each relationship includes a source entity, a target entity, and descriptive information about their connection. When duplicate relationships are found between the same entities, they are merged by combining their descriptions to create a more complete understanding of the connection.

- A community (cluster) is a set of related entities and relationships identified through hierarchical community detection, using the Leiden Algorithm. Communities create a structured way to understand different levels of granularity within the knowledge graph, from broad overviews at the top level to detailed local clusters at lower levels.

However, creating knowledge graphs has been a resource-intensive process, relying on manual curation by domain experts or the conversion of existing structured data from relational databases. Several formal representations have been established to structure the knowledge.

- Graph-based Representations: Knowledge bases are often structured as knowledge graphs, where entities are nodes, relations are edges, and attributes are node or edge properties.

- Ontologies: These define a schema of classes (e.g., "Victim", "Camp") and their relationships, often using standards like OWL (Web Ontology Language). Ontologies provide a hierarchical structure for reasoning, e.g., inferring that all concentration camps are types of locations.

- RDF Triples: Based on the Resource Description Framework, facts are stored as subject-predicate-object triples (e.g., `Anne Frank, wasBornIn, Frankfurt`). RDF enables interoperability across systems, crucial for linking Holocaust archives globally.

In the context of knowledge graphs, formal semantics and ontologies play a crucial role in enabling clear explanations and consistent data organisation. Formal semantics refers to the use of well-defined rules to represent and reason about information, and knowledge graphs rely heavily on it, meaning they use predefined structures (ontology) to define the types of entities and relationships. Ontologies consist of:

- Classes define the types of entities in the knowledge graph, such as "Person", "Place" or "Event".

- Properties describe the attributes of entities and the relationships between them. For example, a "Person" class might have properties like "Name", "Age", and "Address".

- Instances: The actual data points in the knowledge graph represent specific classes.

Knowledge graphs can use ontologies to organise data in a consistent and structured manner, making it easier to analyse.

2.4.7.2 Domain-specific Knowledge Bases

Domain-specific knowledge bases (DSKBs) represent structured information tailored to particular fields or contexts, distinguishing general-purpose KBs such as DBpedia

and Wikidata by incorporating specialised ontologies, entities, and relations that capture the unique semantics and terminologies inherent to a given domain, such as healthcare, engineering, or historical studies (Konys and Drążek, 2020). The DSKBs structure domain-relevant facts as graphs, where entities are represented as nodes and their relationships as edges. This structure enables advanced querying, inference, and integration with machine learning models to power decision-making and knowledge discovery.

A recent survey (Z. Zheng et al., 2020) on the use of LLMs in knowledge graph applications evaluates knowledge graph completion as a fundamental task. Further, some studies have explored (Z. Zheng et al., 2020; Xie et al., 2022) the use of ChatGPT on a link prediction task in the knowledge graph and evaluated its effectiveness. (Z. Zheng et al., 2020) discussed the incorporation of structural information from knowledge graphs into LLMs to achieve structural-aware reasoning. Integrating domain-specific knowledge bases into LLMs, either parametrically or via retrieval augmentation, has been shown to significantly enhance performance on specialised tasks (Di Pasquale and Represa, 2024). Furthermore, comprehensive studies have reasoned about the practical applicability of LLMs combined with knowledge bases. These applications, which include question-answering over customised graphs, demonstrate enhanced reliability for specialised tasks such as medical diagnosis and legal reasoning(S. Ji et al., 2022). Similarly, in the realm of digital humanities, domain-specific knowledge bases can aggregate archival data from historical events like the Holocaust, extracting and linking entities such as survivor names, camp locations, and deportation timelines from unstructured texts to reconstruct narratives and preserve collective memory.

2.4.7.3 Graph Representation Formats

Knowledge graphs can be represented using different formal models, each with distinct characteristics, strengths, and use cases. The choice of these formats affects how knowledge is stored, queried, and reasoned over. The Resource Description

Framework (RDF) and property graphs are considered the most common types of graph formats used for making informed design decisions when constructing knowledge graphs from narrative texts.

Resource Description Framework (RDF)

RDF standards in the form of subject, predicate and object triples, which are especially designed for querying the Semantic Web. RDF enables data integration and reasoning across distributed, heterogeneous sources. The fundamental unit of RDF is the triple, where each triple represents an individual statement about a resource: the subject is the entity being described, the predicate is the property or relationship type, and the object is the value or related entity. All subjects and predicates in the RDF are identified by Uniform Resource Identifiers (URIs), which provide global identification and allow seamless integration of data from different sources. Furthermore, RDF can be extended with vocabularies that define classes, properties, hierarchies, and constraints, providing a semantic layer over the basic triple structure. RDF Schema (RDFS) provides basic ontological primitives, including class definitions, subclass relationships, and property domain/range specifications. To provide richer expressivity for complex ontological modelling, the Web Ontology Language (OWL) extends basic RDF schemas. OWL allows for describing property characteristics such as transitivity (if A acquired B and B acquired C, then A indirectly controls C), symmetry (if A partners with B, then B partners with A), and functionality (an organisation can have only one CEO at a time). Furthermore, OWL enables cardinality restrictions and class expressions using logical operators that support automated reasoning.

RDF schemas offer several significant advantages for knowledge graph construction from narrative texts. Its standardisation as a W3C recommendation means well-defined specifications and broad industry support. The URI-based identification enables the integration of different resources, such as bases such as DBpedia, Wikidata, and Schema.org, to extract knowledge from narrative texts. The integration with OWL supports complex reasoning and enables automated

inference of new facts and validation of consistency. These characteristics make RDF suitable when building knowledge graphs intended for publication as linked open data, integration with existing Semantic Web resources, or applications requiring formal reasoning and ontological rigour.

Property Graphs

Property graphs are a more flexible graph model which is widely used in modern graph databases such as Neo4j, Amazon Neptune, and JanusGraph. Compared with the RDF schemas, property graphs allow both nodes and edges to carry key-value properties, providing greater modelling flexibility. A property graph contains four fundamental elements. Nodes represent entities, each with a unique identifier, one or more categorical labels, and a set of properties. Edges describe relationships between nodes, each containing a unique identifier, a single type label, references to source and target nodes, and their own properties. Labels provide semantic categorisation for both nodes and edges. Properties store data as key-value pairs on either nodes or edges. This structure aligns naturally with how developers and domain experts conceptualise graph data.

The property graph model provides several advantages over RDF's triple-only approach. Most significantly, edges can directly carry properties without any reification. In RDF, representing this same information would require creating an additional resource and connecting it via multiple triples, which results in a more complex structure. Property graphs make this natural, where the relationship edge itself carries all relevant metadata about that specific relationship instance. Additionally, it supports multiple labels per node, allowing entities to belong to several categories simultaneously, and enables more natural modelling with real-world entities that do not fit neatly into single categories. The schema-optional nature of property graphs provides flexibility during development and evolution. Nodes and edges can be created with initial properties, and additional properties can be added later without requiring schema modifications or data migration.

Property graphs perform well in several scenarios relevant to knowledge extrac-

tion from narrative texts. When working within a self-contained system organisation where global URI-based integration is not required, property graphs provide simpler schemas without the overhead of URI management. The ability to attach rich metadata directly to relationships is invaluable when extracting knowledge from texts, as relationships often carry temporal information, confidence scores, source document references, and contextual details. Query performance for traversal-intensive operations, such as finding all entities within N relationship hops or computing centrality measures, is better in property graphs, which are optimised for these operations. The schema-optional nature accelerates iterative development, allowing the knowledge graph structure to evolve as more texts are processed and new patterns are discovered. The intuitive Cypher (query language) syntax reduces the learning curve for analysts and domain experts who need to query the knowledge graph but may not have deep technical backgrounds. However, property graphs have limitations compared to RDF. There is no universal standard for different graph databases, which implement slightly different variations of the property graph model. Moreover, the lack of formal ontology languages integrated into the core model results in having custom implementations for automated reasoning and consistency checking. Integration with external knowledge bases requires custom mapping logic rather than the automatic URI-based linking that RDF provides.

The choice between RDF and property graphs depends on the implications for knowledge graph design, querying capabilities, scalability, and long-term maintainability. However, the decision cannot be made in abstract terms but must be grounded in specific project requirements, technical limitations, and intended use cases. This section examines a domain-specific narrative knowledge graph and demonstrates the reasoning that leads to selecting RDF as the appropriate representation format for this domain. The Holocaust itself presents several characteristics that help to decide the choice between representation formats. The selected survivor testimonies represent a huge area of knowledge extending to the geographical locations, military involvements, family relationships, etc. The knowl-

edge is historically grounded and legally significant, intended for documentation, preservation, and scholarly research. The source material is testimony that may be shared across multiple archives and research institutions. Named entities (people, places, organisations, camps) reference real-world individuals and locations that may be documented in multiple external knowledge bases and historical archives. Relationships are predominantly temporal and familial, with precise dating often uncertain ("before the war" rather than specific dates) and relationships that change over time (locations where people lived at different periods and employment roles held for varying durations). Based on the given factors below, the RDF format is used for knowledge representation for our use case.

- The complex nature and evidence preservation: In a relational or property graph representation, attaching evidence to relationships requires either heavy workarounds or acceptance of information loss. The RDF reification provides a clean, standardised way to represent "this fact came from this text, extracted from this source document on this date", while requiring additional triples. The `meta:evidence` property on each sentence directly links facts to their textual grounding for historical research. For example, if we consider the sentence, *My mother's mother, Freida Borschevskaya, nee Rutenberg, was born in Romny* when representing this sentence in a property graph, one might represent Freida's birthplace by an edge labelled 'born_in' connecting her to Romny. The problem with this model is that it fails to distinguish between the semantic claim itself (that Freida was born in Romny) and the epistemological justification for that claim (the historical record or testimony). In RDF, reification separates these cleanly, creating a Statement resource that explicitly characterises the claim's status, source, and supporting evidence. This separation allows for distinguishing the fact (Freida was born in Romny) from the given sentence.
- Possibility of integration with Semantic Web Archives and Linked Data Infrastructure: Many historical archives and institutional repositories tend

to publish their data using RDF and linked data standards, enabling seamless integration of extracted knowledge with external resources such as DBpedia, Wikidata, etc. The URI-based identification method in RDF defines natural and unambiguous entities. If the knowledge graph were represented in a property graph, establishing those connections would require custom mapping logic outside the graph itself, making integration fragile and non-standard.

- Long-Term Preservation and Standards Compliance: RDF enables the preservation of historical and biographical knowledge for long-term use, spanning decades or centuries. RDF's standardisation by W3C provides assurance that your knowledge graph can be read, validated, and processed using standards-compliant tools far into the future. The Turtle serialisation format is human-readable and language-independent, supporting archival best practices. The SPARQL query language is standardised and documented, ensuring that queries written today will work with tools and systems developed in future years. For institutional knowledge that must persist across organisational and technical changes, RDF's standards-based approach provides stronger guarantees. As another concern, formal validation is required to ensure the logical consistency of structured data, which often contains inherent temporal and relational constraints. For instance, a person's date of death must logically follow their birth date, a child's birth must occur after their parents', and an employment period must have an end date that follows its start date. To enforce these rules, an ontology language like OWL is used to define formal logical axioms and class structures on top of the RDF data, where a reasoner automatically infers new information and identifies inconsistencies within the data, flagging them as validation errors. However, property graphs lack integrated, declarative validation mechanisms where logical rules are implemented through custom application code, which is harder to maintain.

RDF is the appropriate representation format, not despite its complexity but because of its features that address exactly your domain's requirements. The

standardisation, provenance preservation, ontological expressivity, and temporal reasoning capabilities that RDF provides are needed for analysis of historical narratives. While property graphs would be more suitable for different applications (real-time social network analysis, recommendation systems, and transaction networks), to analyse the historical narratives, source verification, institutional archiving, and semantic integration with broader linked data infrastructure are precisely those where RDF's strengths matter most. The choice of RDF for narrative analysis is therefore well-justified on technical and domain-specific grounds.

2.5 Technology and Historical Oral Narratives

With the emergence of neural networks, a new paradigm in computer science has been established, allowing researchers and practitioners to leverage these advanced technologies to address complex and evolving computational challenges (Dean, 2022). In parallel, the field of digital humanities continues to evolve, increasingly integrating neural network models for information extraction in different forms of cultural and historical data, such as oral testimonies, manuscripts, and archival records (Povey, Kingsbury, et al., 2005). This fusion of disciplines enhances data preservation, interpretation, and analysis, enabling deeper insights into historical narratives(Picheny, Zoltán Tüske, et al., 2019). As digitalising the Holocaust resources provides preservation, accessibility, and research opportunities, it also introduces significant challenges in handling the complexities of digitising unique documents (J. V. Psutka, Pražák, and Vaněk, 2021).

While LLMs perform adequately on straightforward sentiment tasks, recent evidence confirms they continue to lag behind fine-tuned encoder models such as BERT and RoBERTa on complex affective and semantic tasks, including emotion cause analysis, emotion recognition in conversation, and aspect-based sentiment analysis (Yiqun Zhang et al., 2026; Lian et al., 2025). This gap is particularly pronounced in domain-specific settings, where general-purpose LLMs using prompt

engineering achieve around 50% accuracy on multi-class emotion recognition compared to over 80% for fine-tuned models, and where even GPT-4 performs significantly below human level on semantic understanding benchmarks (Chang et al., 2024). Furthermore, integrating neural networks into digital humanities workflows enables the creation of interactive and accessible archives, promoting broader public engagement and ensuring that the lessons of the Holocaust remain relevant for future generations.

Recent research highlights that challenges exist in digitising Holocaust testimonies because of the personal, unstructured, and emotionally charged content (Kovács, 2018). Kovács notes that these materials often include fragmented texts and multilingual content, which present significant obstacles for digitisation and make standard text processing insufficient. As a result, considering today's technological landscape, scholars are increasingly turning to advanced computational tools such as natural language processing, automated text analysis, and artificial intelligence to extract meaningful patterns and insights from these rich but challenging sources (Fan and Presner, 2022). This section discusses the various technological efforts undertaken to digitise Holocaust testimonies.

2.5.1 Early Stages of Analysing Holocaust Testimonies

In the late 20th century, significant efforts were made to record the stories of Holocaust survivors to preserve their memories for future generations (Greenspan et al., 2014). The main reason for the digitalisation of the survivors' stories was that, with time, the community had realised the survivors were ageing and would not be able to share their firsthand accounts indefinitely (Greenspan et al., 2014). Over time, the necessity of converting oral testimonies into transcripts has increased to preserve and improve accessibility of these narratives, facilitate research and address ethical and legal considerations, making them a valuable resource for scholars, educators and the public (Bergen, 2019).

However, transcribing these testimonies is a complex and time-consuming

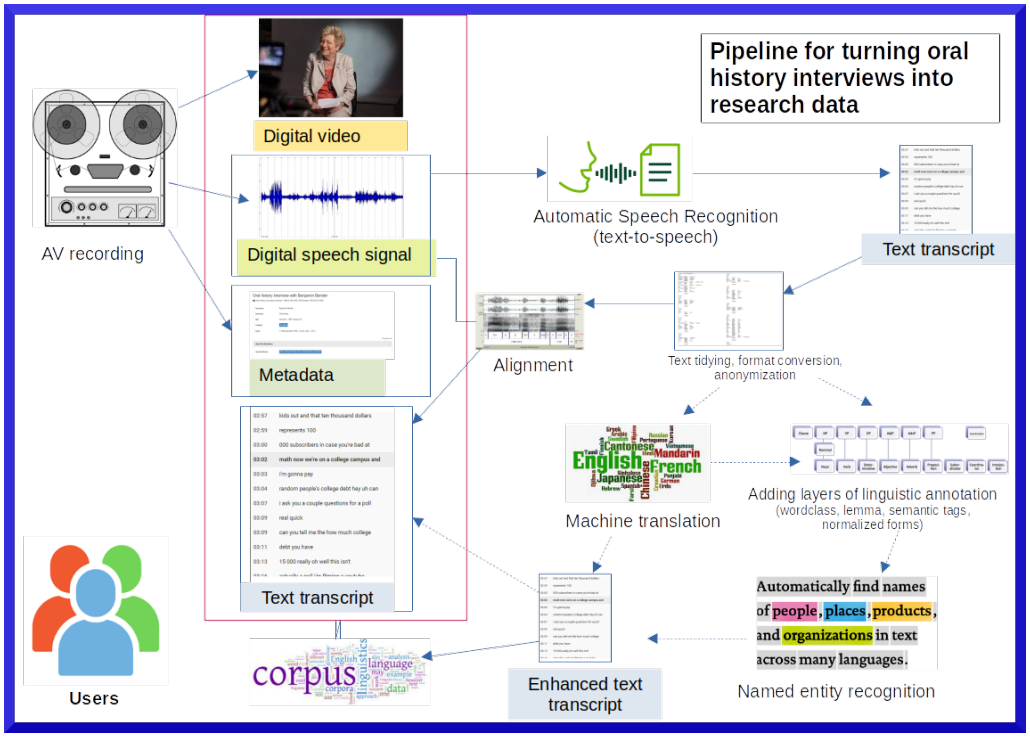

Figure 2.9: Overview of the Digitalisation of Holocaust Testimonies: Workshop Presentation by CLARIN and the European Holocaust Research Infrastructure (Wynne, 2023)

method; therefore, with time, the necessity of a computational solution to extract information on oral testimonies using computational techniques is rising (Pessanha and Salah, 2021). Moreover, institutes such as USHMM [1], Wiener Library [2], and USC Shoah Foundation [3] have collected thousands of survivor testimonies in multiple languages. However, once transcribed, oral testimonies become text-searchable, allowing researchers, educators, and students to explore more in testimonial archives easily.

As an initiative for transcribing testimonies using computational approaches, the NSF-sponsored MALACH project was developed to automate searching oral spoken archives (Byrne et al., 2004). Since the technology is at its developing

[1] https://www.ushmm.org/remember/holocaust-reflections-testimonies
[2] https://www.testifyingtothetruth.co.uk/viewer/
[3] https://vha.usc.edu/home

stages, speech recognition technology was speaker-adaptive processing, sometimes combined with more sophisticated techniques such as MMI/MPE training (Povey and Woodland, 2002) or (by the end of the project) fmpeE processing (Povey, Kingsbury, et al., 2005). However, in 2017, researchers integrated the deep learning features and improved recognition accuracy (Picheny, Zoltán Tüske, et al., 2019). Multilingualism is a massive challenge in oral testimonies. As a result, speech recognition has adapted to other languages, such as Czech (J. V. Psutka, Pražák, and Vaněk, 2021).

2.5.2 Holocaust and Latest studies

At the initial stage of research on Holocaust testimonies, the availability of computational resources was limited, particularly in terms of accessible and standardised digital corpora. Most testimonies were digitised within institutional archives such as museums, requiring formal permission for access, and lacked interoperability across collections. Furthermore, there was no centralised platform to support computational research or to disseminate findings within a dedicated scholarly community. As noted in through this research, organisations such as CLARIN-EU and European Holocaust Research Infrastructure (EHRI) have begun to address these limitations by promoting corpus creation and reproducible workflows for Holocaust testimony analysis Wilson, 2023.

In recent years, domain-specific datasets and computational methods have evolved significantly and increased. For instance, the EHRI-NER corpus provides a multilingual annotated dataset tailored to Holocaust-related texts, enabling the development of specialised Named Entity Recognition (NER) models for entities such as camps, ghettos, and persons (Dermentzi and Scheithauer, 2024). Similarly, the USHMM oral testimonies dataset has been used for tasks including text classification, span categorisation, and NER, demonstrating the growing availability of machine-readable Holocaust testimony data Mattingly, 2025. Post-2022 research, particularly within venues such as the Holocaust Testimony Research Workshop

(HTRES) (**Anuradha**, Wynne, et al., 2024), demonstrates a growing application of advanced NLP techniques to Holocaust testimonies. For instance, Ifergan et al. (2024) apply BERTopic, a transformer-based topic modelling approach, to uncover latent thematic structures in survivor narratives. Their work highlights how contextual embeddings can capture nuanced narrative patterns and variations across testimonies, moving beyond traditional bag-of-words models. Overall, recent developments mark a clear transition from manual and rule-based approaches toward context-aware, scalable, and semantically rich NLP methods. The increasing availability of curated datasets, including recent contributions published Jaff (2025), further strengthens the foundation for computational analysis.

2.6 Identified Research Gaps and Reflection

Despite significant advances in both digital humanities scholarship and computational linguistics, the intersection of these fields in the context of Holocaust testimony analysis remains yet to be explored. This section presents the critical gaps identified throughout the preceding literature review and articulates how these gaps motivate the methodological and empirical contributions in the next chapters.

2.6.1 Limited computational research on oral narratives and Holocaust studies

While Holocaust testimonies have been extensively studied from historical, literary, and psychological perspectives (Langer, 1993), their analysis using computational methods remains limited. The majority of existing computational work in digital humanities has focused on historical newspapers (Neudecker, 2016; Ruokolainen and Kettunen, 2018), archival documents (Poso et al., 2023), and historical books (Hosseini et al., 2021; Manjavacas and Fonteyn, 2021), with oral testimonies receiving comparatively minimal attention. This gap has created space for institutions such as the USC Shoah Foundation, USHMM, and the Wiener

Library to digitise testimony transcripts, representing an underexploited corpus for computational analysis (Brazzo and Speck, 2018).

The limited research that does exist on Holocaust oral histories has primarily employed early-stage speech recognition technologies (Byrne et al., 2004; Picheny, Zoltán Tüske, et al., 2019) or basic keyword search functionalities, without advancing to more sophisticated NLP tasks such as domain-specific named entity recognition, relationship extraction, or semantic knowledge graph construction. As a result, the potential of modern transformer-based architectures and large language models to extract structured historical knowledge from unstructured testimony narratives remains largely unexplored (Blanke, Bryant, and Hedges, 2020).

2.6.2 Lack of domain-adapted language models for oral narratives

Pre-trained language models have demonstrated state-of-the-art performance across numerous NLP tasks, yet their application areas have a fundamental domain adaptation problem. Standard models such as BERT, RoBERTa, and GPT are trained on general web corpora that contain minimal domain-specific terminologies and limited representation of the multilingual code-mixing, Yiddish terminology, and euphemistic language patterns characteristic of survivor narratives (Ionescu and Mitroiu, 2023). While historical language models have been developed for early modern English (Manjavacas and Fonteyn, 2021), 18th-century texts (Hosseini et al., 2021), and Italian historical documents (Palmero Aprosio, Menini, and Tonelli, 2022), no equivalent effort has been undertaken for oral testimonies or survivor stories of the Holocaust. This absence is problematic given the unique linguistic features of the domain: fragmented sentences resulting from trauma (Laub and Auerhahn, 2017), indirect speech acts to manage emotional weight (Kraft, 2004), and domain-specific hyponymy where survivors may use general or specific terms depending on their psychological state during testimony (Bailey et al., 2020). The absence of a generalisable, domain-adaptable pre-trained model creates a critical

performance gap: existing NLP systems consistently underperform on Holocaust-specific tasks, resulting in the misidentification of entities, a misinterpretation of context, and an inability to capture the semantic richness of survivor speech.

The complex linguistic patterns in oral language make it more challenging to analyse than written language. Informal speech patterns, such as interruptions and overlapping dialogue, are common in conversations. Oral language is often unstructured, consisting of fragmented sentences, pauses, repetitions, and interruptions. Unlike written text, which follows grammatical rules and structured syntax, spoken language is context-dependent, making it difficult for NLP models to parse and analyse accurately. Moreover, oral testimonies frequently contain ambiguous phrases, filler words (e.g., "um," "uh") and background noise (e.g., coughing, overlapping speech). These elements introduce acoustic and textual inconsistencies that significantly degrade the accuracy of speech-to-text models. When survivors describe their memories, their speech often carries emotional weight, conveyed through tone, pitch, volume, and pacing. Research in affective computing and speech emotion recognition suggests that prosodic features (such as intonation and stress patterns) are essential in conveying emotions. Yet, they remain challenging for modern NLP models to interpret.

When considering the Holocaust as one historical event, spoken language evolves, incorporating new slang, idioms, and expressions. In narratives, survivors use terms that were common in wartime ghettos, concentration camps, or resistance movements, many of which do not exist in modern vocabulary. State-of-the-art NLP models trained on general datasets may struggle to understand domain-specific, historically and culturally specific terminology, limiting their applicability in Holocaust testimony analysis.

2.6.3 Absence of annotated and standardised datasets for Holocaust-specific NLP tasks

A critical bottleneck in developing computational tools for Holocaust testimony analysis is the near-total absence of publicly available, expertly annotated datasets. Domain-specific NER systems require labelled training data that captures the unique entity categories relevant to survivors' oral narratives(Ehrmann et al., 2023). Similarly, relationship extraction models depend on annotated examples of semantic relations given in the testimonies. The absence of annotated and standardised datasets has multiple consequences. First, researchers are unable to fine-tune pre-trained models effectively for domain-specific tasks. Second, the lack of standardised annotation schemas prevents comparative evaluation across different computational approaches (De Toni et al., 2022).

Standardisation of Holocaust testimonies is a challenging task when digitising as a historical resource. Since different organisations use different objectives and policies to collect and preserve, bringing all of them to common ground is difficult. Each organisation operates with its mission, target audience, and resources, leading to differences in how testimonies are recorded, transcribed, archived, and shared. Some organisations may focus on preserving raw, unedited testimonies to maintain authenticity, while others make them more accessible for research and educational purposes.

According to existing collections and archives, two structures are available: narrative style and the interview format. In narrative style, the survivor explains what they have witnessed in the Holocaust. This format allows survivors to share their stories freely, often detailing their lives before, during, and after the Holocaust without direct supervision or interruption. The interview format involves a structured approach, where an interviewer poses a predefined set of questions to the survivor. This guided format ensures that specific events of the survivor's experience are covered, including their life before the Holocaust, their experiences during the events, and their life afterwards. However, some survivors may not answer

specific questions, particularly those that are personal or sensitive, as revisiting such memories can re-traumatise them and evoke painful emotions.

Both structures have their strengths and limitations. The narrative style provides a more organic and personal account, and the interview form delivers a structured and comprehensive overview of the survivor's life experience. However, this diversity introduces additional complexities when processing documents using computational approaches. The unstructured nature of narrative-style testimonies makes it challenging to automatically extract and categorise metadata (e.g., dates, locations, events) separately from emotional or subjective content. In contrast, while interview formats are more structured, they still contain inconsistencies, such as unanswered questions and variations from the script, because survivors recall their memories, and memories do not appear orderly. While both formats enrich the understanding of the Holocaust, their structural differences pose significant challenges for automated processing, particularly in distinguishing metadata from emotional data and ensuring the accuracy of digital transcripts. To address these challenges, advanced computational techniques are required in the future with careful consideration of the unique features in each format.

2.6.3.1 The multilingual and code-mixing nature of the Holocaust Testimonies

Code-mixing is a common scenario in Holocaust testimonies where words, phrases and grammatical structures from different languages blend within a single discourse. Code mixing can happen due to historical, geographical, and emotional aspects associated with the Holocaust testimonies, making the automated identification, transcription, and translation of testimonies particularly challenging. Holocaust survivors came from diverse linguistic and cultural backgrounds from multiple European territories, each with its own language. As a result, survivors frequently switch between languages when describing locations, events, and cultural incidents that happened during the Holocaust. For example, a survivor could be deported

across different occupied regions or camps such as Auschwitz (German) or Oświęcim (Polish) or use local terms from their native Yiddish or Hebrew. Moreover, the emotional intensity and trauma experienced by Holocaust survivors are key reasons that often lead to code-mixing in the testimonies. As survivors recount their memories, they may switch from a single language to their native language, specifically when describing deeply distressing or traumatic events. This transformation often occurs because extreme emotions, such as fear, pain, or sorrow, can trigger a natural return to the language most closely tied to their identity and early experiences. Such code-mixing reflects their psychological state and adds cultural and emotional depth to Holocaust narratives. Further, Holocaust survivors combine grammatical structures from different languages in their speech, which is more than simple word mixing, making automated processing even more difficult. However, state-of-the-art models struggle to recognise code-mixing due to the limited availability of high-quality annotated datasets compared to monolingual data. The complex linguistic nature of code-mixing, which involves mixing grammar and syntax from different languages and diverse patterns. Although code-mixing is prevalent in Holocaust-related documents, some languages used in these texts are considered low-resource, making it difficult for computational approaches to process information in these languages accurately.

2.6.4 Limited integration of knowledge representation and visualisation

Although knowledge graphs and geospatial visualisations have proven valuable for historical research (Hogan et al., 2021; Knowles, Cole, and Giordano, 2014), their application to Holocaust testimonies remains fragmentary. Existing digital Holocaust archives provide keyword search and metadata filtering, but they do not offer structured semantic query capabilities that would allow researchers to trace complex relational patterns, such as identifying all survivors who passed through a specific transit camp, mapping family networks disrupted by deportation, or

reconstructing temporal sequences of events across multiple testimonies (Kovács, 2018).

The construction of domain-specific knowledge bases from unstructured testimony text requires the integration of NER, relationship extraction, temporal reasoning, and entity resolution. Furthermore, the visualisation of such knowledge must be designed with sensitivity to the ethical responsibilities of representing survivor experiences, avoiding reductionist or decontextualised representations that distort historical understanding.

As the gaps above indicate, the necessary computational tools and methodologies for analysing oral histories, such as Holocaust testimonies, exist in theory. However, their adaptation, evaluation, and ethical deployment in this specific context have not been achieved. The proposed research addresses these identified gaps through a multi-faceted approach that integrates:

1. **Domain-adapted language modelling**: Development of fine-tuning or pre-trained models on Holocaust testimony corpora to capture domain-specific vocabulary, multilingual patterns, and the distinctive linguistic features of survivor narratives.
2. **Annotation framework development**: Creation of expert-annotated datasets for Holocaust-specific NER and relationship extraction, establishing benchmark resources for future computational research in this domain.
3. **Knowledge base construction and visualisation**: Integration of NLP outputs into structured semantic representations that support complex historical queries while maintaining fidelity to the narrative complexity of survivor accounts.
4. **Ethically grounded computational methodology**: Analysis of ethical dimensions of applying automated computational methodologies to survivor testimony, ensuring that computational methods serve rather than distort historical understanding and collective memory.

By addressing these interconnected gaps, this research contributes not only technical innovations in domain-specific NLP but also methodological frameworks for the responsible application of computational methods to sensitive historical materials. The trained adapter transforms user queries through four sequential stages: query processing, similarity search, context augmentation, and response generation. This implementation demonstrates the practical applicability of the approach combined with the RAG pipeline. The following chapters detail the specific approaches, experiments, and evaluations undertaken to realise these contributions.

2.7 Chapter Summary

Due to the interdisciplinary nature of the research and the lack of existing studies that directly address the intersection of historical significance in oral testimonies, linguistic patterns in spoken text, and the adaptability of computational tools, this chapter surveys relevant literature from other similar studies conducted within the last decade. While there has been progress in automatic information extraction techniques on general-purpose applications, only a limited number of studies have been conducted with the unique challenges posed by sensitive, trauma-informed and spoken text, such as Holocaust testimonies. A further limitation is the absence of a centralised infrastructure tailored for historical domain-specific pipelines, which hinders consistent, reusable, and ethically informed information extraction practices. Therefore, this thesis explores oral language processing and NLP methods in historical contexts to identify these critical gaps and establish the need for the proposed infrastructure. While NLP and LLMs hold great potential for analysing spoken language, significant challenges remain, particularly in the context of complex and sensitive testimonies like those of Holocaust survivors. Therefore, the next chapters will discuss the annotated dataset and different machine learning techniques for information extraction and integration through a unified framework for knowledge representation.

Part I

Corpus, Extraction and Representation

Chapter 3

Holocaust Testimonies as Oral Historical Corpora

Everybody, every human being has the obligation to contribute somehow to this world.

Edith Carter (Survivor)

To a large extent and for decades, the Holocaust testimonies have been scattered and virtually inaccessible. Most of these historical documents have been digitised with the advancements in digital humanities, which allow computational methods to process them and extract information automatically. Through Holocaust testimonies, survivors bear witness by narrating their lived experiences, intending to make the audience knowledgeable. The proposed dataset represents the labelled corpus of Holocaust survivor testimonies that reflect deeply personal accounts of a historical event, shaped by memory, trauma, and identity. Unlike conventional textual corpora, these testimonies consist of oral language that exhibits significant emotional depth, unstructured narrative styles, and frequent temporal disjunctions. This chapter analyses the nature, structure, and types of information within Holocaust narratives and investigates the procedures and challenges involved in annotating survivor testimonies as the gold dataset for different NLP tasks, which would be further discussed in the next chapters.

3.1 Nature of the Holocaust Testimonies

Holocaust testimonies are emotionally complex texts that are often conveyed through unstructured and deeply personal stories. As a result, when developing computational tools, additional attention is required when analysing the information embedded within survivor narratives. The incomplete nature of traumatic memories, combined with the personal experiences, requires critical evaluation and analytical approaches that are able to analyse both factual and emotional content. According to this chapter, the information contained in Holocaust testimonies is primarily categorised into four types: biographical, temporal, geographical and emotional data.

- **Biographical Information:** Biographical information includes the survivor's name, date and place of birth, nationality, residential history, religious/ethnic identity, pre-war life and occupation. Biographical information provides more insights about the family backgrounds and structures, including the names of relatives, pre-war social networks, and community affiliations. Moreover, some survivors testify about post-war biographical information, such as migration patterns, name changes, and efforts at rebuilding lives, to understand the long-term impact of genocide on survivors' identities.
- **Temporal information:**To understand the timeline of a survivor's experience and for reconstructing historical events, temporal factors such as dates, durations, chronological sequences and historical periods were extracted from Holocaust testimonies. The temporal information accounts for both chronological and temporal references, including relative time expressions ('weeks later', 'before the deportation') and contextual temporal cues that require historical knowledge to interpret accurately.
- **Geographical Information:** Geographical information provides significant factors of the Holocaust testimonies. Survivors reference a wide range of spatial entities, including cities, towns, villages, countries, rivers, forests, streets, and administrative regions. More importantly, survivors highlight

deportation routes, forced marches, transfers between camps, hiding locations, and eventual post-war displacement. Extracting this spatial data allows researchers to reconstruct the geographical trajectories of persecution, map the reach of the Nazi apparatus, and identify patterns in survivor migration.

- **Emotional data:** As the personal narratives of the Holocaust, testimonies consist of expression, description, and trauma experienced by the survivors. When describing specific events such as deportation, separation from family, and liberation, survivors verbalise emotions explicitly (e.g., 'I was terrified' and 'I felt sad'). Beyond these direct emotional statements, testimonies convey emotional effects through linguistic markers, including use of silence or omission paralinguistic features captured in audio-visual testimonies such as pauses, tone, and body language. Furthermore, the emotional data in testimonies includes not only fear, grief, and despair but also moments of resilience, solidarity, resistance, and hope, which together form a more complete picture of the survivor experience.

3.1.1 Structure of Holocaust Testimonies

The growing digitisation of Holocaust testimony archives has created space for research in digital humanities and computer science. However, their use as data depends on the format and method of collection. The structure of a testimony directly affects its linguistic properties, narrative coherence, and the types of information that can be extracted. The most common structural types are discussed below.

Narrative-style testimonies

The narrative-style testimonies are unstructured, in which survivors are encouraged to recount their experiences in their own words without guidance or interruption from a third-party. Narrative-type oral testimonies are a form of qualitative, first-person, and autobiographical accounts. Within Holocaust studies, oral testimonies serve as first-hand accounts that provide insights about

the Holocaust, encompassing emotional, historical, and socio-cultural information. Narrative testimonies allow survivors to share their experiences in their own words, choosing what to emphasise and how to express them. As a result, narrative-type testimonies include diverse linguistic patterns, code-switching, and emotionally charged language, which make them complex to analyse computationally. Additionally, survivors may recall events out of chronological order, reflecting the fragmented and incomplete memories. This non-linear structure further complicates automated processing but provides deeper insight into personal and historical trauma. Examples of sources of this type include large-scale oral history archives such as the USC Shoah Foundation Visual History Archive. Similarly, the Fortunoff Video Archive for Holocaust Testimonies at Yale University and the Wiener Holocaust Library preserve narrative-style testimonies that emphasise uninterrupted, first-person accounts.

Interview-type testimonies

Interview-type (questionnaire-based) testimonies differ from narrative-style ones, as they are structured around a question and answer format, guided by an interviewer. The interview-type testimonies follow a semi-structured and structured format, designed to extract specific information such as historically specific events, geographic locations, personal identities, and experiences related to deportations, persecution, and survival. Because of that, the nature of information across multiple survivors is comparable while preserving each narrative's personal dimension. The dialogic nature of interview-type testimonies introduces distinct linguistic features. For example, the presence of both interviewer and interviewee introduces turn-taking, interruptions, clarification requests, and topic shifts. These features result in a conversational style that is often more fragmented and context-dependent than monologic narratives. Examples of interview-type testimonies can be found in structured oral history collections such as those provided by the United States Holocaust Memorial Museum, where interviews are conducted using guided questionnaires to capture specific historical details.

Table 3.1: Statistical overview of the testimony corpus.

Property	**Value**
Primary source archives	Wiener Holocaust Library, United States Holocaust Memorial Museum, Fortunoff Video Archive, Centropa Archive
Testimony format	Narrative (N=1367), Interview (N=1564)
Primary language	English
Mean transcript length (tokens)	N=856

3.2 Dataset Description and Corpus Statistics

The corpus assembled for this study consists of more than 3000 Holocaust survivor testimonies drawn from publicly accessible archives. The primary sources include the *Wiener Holocaust Library, UK*, *the United States Holocaust Memorial Museum*, *Fortunoff Video Archive* and the *Centropa testimony collection* . These archives were selected because of the structural and stylistic diversity of real-world Holocaust testimony data. Table 3.1 provides a statistical overview of the corpus. The corpus covers testimonies delivered in the English language but includes different terms in German, French, and Hebrew, etc. According to our analysis, survivors range in age at the time of recording from child survivors of the Holocaust to survivors of concentration camps. However, the whole dataset was not subjected to human annotation due to the time constraints. Only the set of 200 testimonies was selected for human annotation.

3.3 Data Annotation pipeline

At the outset of this research, no annotated dataset existed for Holocaust survivor testimonies that could serve as a gold standard for named entity recognition and relationship extraction tasks. This absence made the construction of such a resource the necessary first step before any computational pipeline could be developed or

evaluated. The annotation strategy was designed in three stages, driven by both the scale of the corpus and the constraints of manual annotation in a domain requiring specialised historical expertise. With over 3,000 testimonies in the corpus, manually annotating the entire collection was neither feasible nor sustainable. The domain-sensitive content makes annotation time-intensive, and each annotator requires not only linguistic competence but also historical knowledge of the Holocaust to label entities and relationships accurately and responsibly. To address this, 200 testimonies were selected for full manual annotation by domain experts, following detailed annotation guidelines developed in collaboration with historians. These 200 testimonies were drawn from three distinct archives in a 70:70:60 ratio: the Wiener Holocaust Library, the United States Holocaust Memorial Museum (USHMM), and the Fortunoff Video Archive.

This manually annotated subset constitutes the gold dataset that underpins all subsequent NLP tasks in this thesis, including named entity recognition, relationship extraction, and knowledge graph construction.

The remaining 2,800 testimonies also required annotation for NER and relationship extraction tasks in order to make the full corpus computationally usable. However, deploying automated annotation methods directly to the larger corpus without prior validation would have introduced uncontrolled errors into a historically sensitive dataset. To determine the most accurate and reliable approach before deploying it at scale, a comparative evaluation was first conducted using the 200 manually annotated testimonies as the reference standard. Both the spaCy rule-based approach and the LLM-based approach were applied independently to this gold subset, and their outputs were measured against the manual annotations. This evaluation provided a principled basis for selecting the most appropriate methodology for annotating the remaining testimonies, balancing accuracy and scalability. This section outlines each stage of the annotation pipeline in detail, beginning with the manual annotation process, followed by the semi-automated spaCy approach, and concluding with the LLM-based approach before presenting

the comparative evaluation that determined which method was applied to the full corpus. The data selection and overall pipeline are illustrated in Figure 3.1.

Table 3.2: Selected list of tags for annotation.

Categories	**Type of Entities**
Domain Specific	GHETTO, SHIPS, CONCENTRATION CAMP, MILITARY RANK, HAPPENING, DISEASE
Location	CITY, RIVER, FOREST, MOUNTAIN, STREET, COUNTRY
Nationality	ETHNICITY, LANGUAGE, NATIONALITY
General Entities	DATE, ORGANISATION, PERSON
Relationships	*taken to, located in, born in (country), born on (date), killed in (place), killed on (date), escape from, escaped to, killed by, worked as, hidden in, married to, transported to, arrived on, left to, lost at, returned to, death of, survived, arrested in, deported to*

Table 3.2 defines the categories and elements to be tagged within the testimonies.

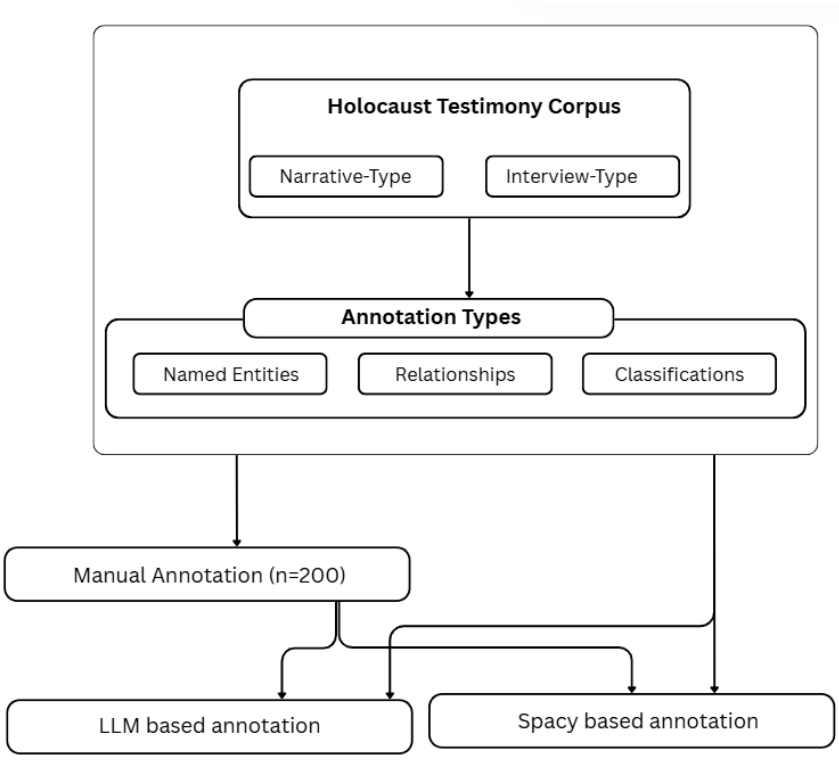

Figure 3.1: Procedure of the data annotation pipeline.

This taxonomy was designed with the support and guidance of historians and domain experts to extract historical, geographical, and personal information from Holocaust testimonies. During the process of defining the annotation schema, several candidate categories were considered and ultimately excluded from the final schema. For instance, religious affiliation and socioeconomic status were initially proposed as separate entity types but were merged under biographical entities due to limited frequency and high annotation ambiguity. The relationship types were selected specifically to support downstream tasks such as knowledge graph construction, focusing on relations that are required to build a short biography of a survivor which is historically informative.

3.3.1 Manual Annotation

The annotation process was conducted in collaboration with domain experts specialised in Holocaust studies to ensure historical accuracy and contextual relevance throughout. Given the linguistic and emotional complexities inherent in survivor testimony, expert guidance was critical at every stage of the process.

The selection of an appropriate annotation tool was important, given the domain-specific complexity of Holocaust testimony annotation and the need for a platform capable of supporting both named entity and relationship annotation within a single interface. Several tools were evaluated prior to selection, including Label Studio, Prodigy, and Doccano. While all three are widely used in NLP annotation tasks, each presented limitations in the context of this research. Label Studio, although highly configurable, required substantial setup for relationship annotation and lacked native support for the simultaneous display of entity and relation labels in a format suitable for non-technical domain experts. Prodigy offered strong integration with spaCy pipelines but operates on a commercial licence, and its active-learning interface was not well suited to the context-dependent annotation required for historically sensitive material. Doccano supported sequence labelling but provided limited functionality for relation annotation between entities spanning

multiple sentences, which is a common requirement in testimony discourse. The UBIAI annotation tool was selected on the basis of four criteria relevant to this domain. First, it provides a unified interface for both NER and relationship annotation, allowing annotators to label entities and the relations between them within the same document view without switching between tools or annotation modes. Second, it supports the import and export of annotations in standard NLP formats, including CoNLL and JSON, ensuring compatibility with downstream model training pipelines. Third, its interface is accessible to domain experts without programming knowledge, which was essential given that the annotation team included historians. Fourth, it supports multi-annotator workflows with built-in inter-annotator agreement tracking, facilitating the quality control process described in the following section.

The annotation guidelines encompassed detailed entity definitions and examples of context-dependent annotations specific to Holocaust narratives. The complete annotation guidelines are attached in Appendix A.1. Inter-annotator agreement was measured on a randomly selected subset of 20% of the annotated testimonies, which were independently annotated by two expert annotators. Cohen's kappa coefficient (κ) was computed at the entity level, yielding a score of average $\kappa = 0.89$, indicating near-perfect agreement in NER. Most disagreements came from pronoun and coreference resolution, as ambiguous wording in the testimonies made it hard for annotators to consistently match pronouns to the right referents. This difficulty is reflected in a moderate inter-annotator agreement for relationship labelling ($\kappa = 0.67$), consistent with the known challenges of coreference annotation in complex narratives. The disagreements that occurred during the annotation process were systematically resolved through meetings with domain experts, and the suitable annotations were used as the final gold standard.

3.3.2 Semi-Automated Annotation

Semi-automated annotation is a hybrid approach that combines human effort with automated algorithms to annotate datasets. Although manual annotation requires more time and effort, semi-automated annotation is prone to errors without human validation. Therefore, we employed a balanced approach by using computational techniques to speed up the labelling process while ensuring accuracy. This approach significantly reduced annotation time by approximately 40-60%. This is achieved by following the post-annotation phase with expert verification, correction, and enrichment of the automatically generated labels. In our annotation pipeline we developed a domain-specific rule-based model using the spaCy library.

Using spaCy for annotating Holocaust testimonies is a rule-based approach to extract meaningful information from historical texts, such as named entities (people, places, and organisations) and custom annotations relevant to the Holocaust. In this process, we have developed regex-based rules and combined them with spaCy's Rule-Based Matcher. Table 3.3 displays the regular expression (regex) patterns created to extract domain-specific named entities, such as `ghetto`. The rule-based patterns were developed iteratively through analysis of a series of testimonies, incorporating variations in terminology, spelling inconsistencies and historical naming conventions. Custom entity rulers were implemented to recognise Holocaust-specific entities, including concentration camps, ghettos, and military groups. Pattern matching was enhanced with contextual rules to distinguish between specific entity mentions and general references.

Table 3.3: Regular expressions designed for spaCy

Entity	**Regex Expression**	**Match**
Street	If name followed by street semantically identical word ([A-Z][a-z]*(strasse|straße|straat)\b or [A-Z][a-z]*(Street|St|Boulevard|Blvd|Avenue|Ave|Place|Pl)(.)*)	Hauptstraße
Ghetto	Search on the lexicon consist Ghetto names or either name followed by ghetto [A-Z]\w+((-|)*[A-Z]\w+)* (g|G)hetto	Anyksciai

Although we utilised spaCy's transformer-based model, which demonstrated the best performance during experimentation, named entity recognition was further refined through the integration of domain-specific regex rules and custom entity matchers. Yet, still, the hybrid model produced incorrect annotations, incorrectly labelling words that matched syntactically because the model lacked contextual or semantic relevance to Holocaust-specific entities.

3.3.3 LLM-based Annotation

With the emergence of LLMs, automatic annotation has become a viable alternative to traditional methods, addressing both the high costs and the accuracy limitations of manual and rule-based annotation methods. By understanding deeper contextual meaning, LLMs demonstrate above average performance in recognising domain-specific entities and relationships. Moreover, recent studies have shown that LLMs have delivered promising results when annotating unstructured text (M. Li et al., 2023). The performance of LLMs depends heavily on the quality and specificity of the prompts provided.

Table 3.4 illustrates the systematic prompt engineering approach incorporating few-shot learning, providing annotated examples from the manual annotation phase that explicitly define the Holocaust-specific entities and contextual guidelines for handling ambiguous cases. For this experiment, the DeepSeek API was employed to automate the annotation process. The prompts were structured to include (1) task definition and objectives; (2) entity taxonomy with detailed descriptions and constraints; (3) 3-5 samples demonstrating correct annotations in various contexts; and (4) output format specifications ensuring consistency with the annotation schema. The LLM-based approach handled contextual ambiguity more effectively than the rule-based method and achieved better recall for rare or complex entities. However, this approach also required careful validation to prevent hallucinated annotations when the model confidently labelled entities or relationships that were not present in the source text. A human-in-the-loop validation step was therefore

Table 3.4: Zero-shot COT Prompt.

Zero-shot COT Prompt-Holocaust Testimonies corpus
Consider the years from 1936 to 1944. You are going to identify name entity tags for holocaust-specific tags. The list of name entity tags should be {list_of_tags}. Each tag is as follows: {tags_meaning}. Now do the below tasks. 1. Try to identify the most suitable Name entity tag for the word 'NAMEENTITY' in the GIVEN SENTENCE based on the below criteria: • Analyse the word in front of the 'NAMEENTITY' tag before you tag. • Understand the complete sentence and try to identify specific factors discussing the word you want to tag. The GIVEN SENTENCE: {sentence}. 2. Return only the GIVEN SENTENCE after assigning the identified tags instead of the word 'NAMEENTITY'. Do not add additional data. Use the following format for the output: "<Updated sentence with correctly identified name entity tags>"

retained for all historically sensitive entities and events.

3.3.4 Comparative Evaluation of Annotation Methods

Before deploying either the spaCy rule-based approach or the LLM-based approach to annotate the remaining 2,800 testimonies, a systematic comparative evaluation was conducted using the 200 manually annotated testimonies as the reference gold standard. Both automated methods were applied independently to the same 200 testimonies, and their outputs were evaluated against the gold annotations using the F1 score. It is important to note that this comparative evaluation was conducted

specifically for the NER task. Both spaCy and the LLM-based approach were evaluated against the gold standard NER annotations, and the results in Table 3.5 reflect NER performance exclusively.

Table 3.5: Comparative evaluation of spaCy vs. LLM-based NER annotation against a gold standard, reporting Precision, Recall, and F1 scores for specific entities.

Entity Category	**Gold**	**spaCy**			**LLM**		
	Count	**P**	**R**	**F1**	**P**	**R**	**F1**
PERSON	475	0.95	0.89	0.92	0.94	0.96	0.95
DATE	324	0.90	0.81	0.85	0.89	0.91	0.90
ORGANISATION	124	0.86	0.75	0.80	0.87	0.89	0.88
CONCENTRATION CAMP	247	0.87	0.76	0.81	0.88	0.90	0.89
GHETTO	134	0.81	0.68	0.74	0.80	0.82	0.81
LOC	356	0.80	0.69	0.74	0.75	0.77	0.76
MILITARY RANK	56	0.75	0.57	0.65	0.78	0.80	0.79
DISEASE	77	0.67	0.51	0.58	0.63	0.65	0.64

Although the LLM-based approach achieved slightly higher overall F1 scores for NER, the decision regarding which method to deploy for annotating the remaining 2,800 testimonies was not based solely on performance. Computational cost was also a significant practical consideration. Deploying an LLM via API for NER annotation across 2,800 testimonies requires substantial financial cost, whereas the spaCy rule-based approach, once developed with domain-specific vocabulary and custom entity rules, operates at negligible computational expense. Given that the spaCy approach achieved competitive F1 scores for well-defined entity categories and that its performance on the full entity set, while lower than the LLM, remained within an acceptable range for large-scale corpus annotation, spaCy was selected as the annotation method for NER across the remaining 2,800 testimonies. All spaCy-based NER annotations were verified and corrected by domain experts to ensure historical accuracy and contextual validity. For relationship extraction, however, a rule-based approach was not viable. As discussed

further in Section 5, the relationship types required by this corpus are expressed through linguistically variable, contextually dependent, and frequently implicit constructions in testimony discourse that rule-based pattern matching cannot reliably capture. Relationships such as *deported to*, *escaped from*, and *hidden in* are rarely expressed through consistent syntactic patterns across testimonies, and their correct identification requires an understanding of narrative context that extends across sentence boundaries, including coreference chains and temporal sequencing. For these reasons, the LLM-based approach was selected exclusively for relationship annotation across the full corpus. Relationship annotations were generated using the DeepSeek API, with prompts structured according to the annotation guidelines developed during the manual annotation phase. All LLM-generated relationship annotations were manually reviewed and validated by domain experts to guard against hallucinated or historically inaccurate relational assignments.

3.4 Challenges when annotating Testimonies

Transcribed Holocaust testimonies pose significant challenges in the process of information analysis due to the linguistic ambiguity, structural variability, and trauma-infused discourse. The above issues arise from the nature of traumatic memory and the conditions under which testimonies were recorded. Survivors frequently use imprecise language when describing traumatic events (Langer, 1993). In testimonies, pronoun-based sentences such as `they took us` which lack clear referents, and vague spatio-temporal references such as `somewhere in the East, around 1944` are frequently visible in testimonies due to memory fragmentation and the effects of extreme trauma. Moreover, survivors mention `the camp` without specifying which one, requiring historians to cross-reference other details. Since these ambiguities complicate NER recognition, an additional coreference resolution layer has to be introduced.

Unlike structured datasets such as government records or standardised surveys,

Holocaust testimonies have significant variability in format, length, and narrative style. The professional background of interviewers, whether psychologists, historians, or documentarians, deeply influences the structure, focus, and content of Holocaust testimonies. The cultural and linguistic background of the testimonies is always influenced by the survivors' native languages and cultural beliefs. Moreover, the purpose of recording context and the time, place, and medium of testimony collection introduces further variability, where official archives follow structured protocols, while family-recorded testimonies may be informal and digressive.

Based on the above-discussed factors, code-switching alternates between multiple languages within a single narrative, reflecting survivors' memories. Survivors naturally switched between languages such as Yiddish, German, Polish, and Hebrew to express concepts that lacked direct translations in their post-war languages; for example, German Nazi terminology such as `Appell`:[roll call] or `Selektion`: [selection]. Moreover, in testimonies, pauses, repetitions, and abrupt topic shifts are common features. These disfluencies serve as meaningful manifestations of traumatic memory, reflecting both the psychological impact of the Holocaust and the cognitive impulses induced when navigating painful narratives.

3.4.1 Annotator Wellbeing and Ethics

Working extensively with Holocaust testimonies exposes annotators to descriptions of violence, loss, persecution, and psychological trauma. Secondary traumatic stress (STS) is a well-documented phenomenon in which individuals who engage with another person's traumatic experience may themselves develop symptoms analogous to post-traumatic stress (Sexton, 2025). This risk is particularly acute in the context of sustained, repeated engagement with large volumes of survivor testimony, as required in the construction of an annotated corpus. To mitigate this risk, several practices were incorporated into our annotation guidelines (Appendix A.1). First, annotators were briefed on the nature of the material prior to beginning work and were given the opportunity to withdraw at any stage. Second, annotation sessions

were time-limited to a maximum of two hours of continuous testimony engagement, with mandatory breaks between sessions. Third, access to psychological support and regular supervision meetings were made available throughout the annotation period. These practices reflect broader guidance on trauma-informed research methodology and are recognised as essential components of responsible digital humanities work involving testimony corpora.

The computational analysis of Holocaust testimonies raises a set of ethical obligations that go beyond standard research ethics requirements and must be explicitly acknowledged. All 200 testimonies used in this study were sourced from publicly accessible archives that have obtained consent from legal representatives for educational and research use. We have contacted the responsible archives and have obtained their consent to use their collection for our research. Furthermore, computational extraction of information from testimonies carries the risk of reducing deeply personal narratives to structured data points, potentially stripping them of the complexity, humanity, and moral weight they encapsulated. Throughout this research, the annotated dataset is understood as a supplement to, not a replacement for, qualitative historical scholarship. No claim is made that NLP-derived outputs constitute historically authoritative interpretations of individual testimonies. The involvement of domain historians at every stage of the annotation process was, in part, a structural response to this risk.

3.5 Chapter Summary

This chapter presented the construction of the first gold corpus of Holocaust survivor testimonies designed for NLP research. It began by examining the unique linguistic and structural properties of oral historical narratives that distinguish this corpus from conventional textual datasets. The annotation strategy was driven by a central practical constraint: with over 3,000 testimonies in the corpus and no prior annotated resource available for this domain, a three-stage

approach was required. Two hundred testimonies were manually annotated by domain experts to produce a gold standard; the remaining 2,800 required automated annotation; and a comparative evaluation against the gold standard was conducted before either automated method was deployed at scale. Three annotation methodologies were systematically evaluated: manual expert annotation, semi-automated rule-based annotation using spaCy, and LLM-based annotation using prompt engineering. Manual annotation ensured historical accuracy but proved time-intensive at scale. The rule-based spaCy approach accelerated the process but struggled with non-standard terminology and contextual ambiguity. The LLM-based approach demonstrated stronger contextual understanding and better recall for complex entities but required rigorous human-in-the-loop validation to prevent hallucinated annotations in historically sensitive contexts. The wellbeing of annotators working with traumatic content was treated as an ethical obligation, with structured protocols implemented to mitigate secondary traumatic stress.

Part II

Information Extraction from Domain-specific Narratives

Chapter 4

Domain-Aware Entity Recognition from Oral Narratives

I have come to realise that my testimony carries so much more weight with them than anything that they have read in a book or see in a movie.

Charles Middleberg (Survivor)

In the field of digital humanities (DH), analysing historical oral narratives requires robust and domain-specific named entity recognition (NER) systems. Most of the oral narratives which belong to the survivors of genocides are linguistically diverse and emotionally sensitive, requiring computational models capable of context-aware text processing. However, previous studies have rarely explored machine learning-based NER frameworks across various domains and their application to oral narratives (Dunstan et al., 2024). Recent work González-Gallardo et al. (2023) presented an application of LLMs for historical documents, demonstrating strong performance in zero-shot learning on domain-specific data after being appropriately fine-tuned. Additionally, informal speech patterns, dialectal variations, code-switching, hesitations, and non-standard grammar are present in oral narratives, complicating the automatic identification and classification of entities using traditional NER models (Sharma and Magar, 2024) (M. Nguyen and Z. Yu, 2021).

This chapter examines domain-specific adaptations of pre-trained and fine-tuned language models applied to Holocaust-related testimony corpora to address the complexities inherent in oral narratives. This approach enables the identification of specialised terminology, entity patterns, and narrative structures within Holocaust testimonies. Through training on specialised datasets, the model learns to recognise the subtle connections between various named entities such as locations, organisations, and events, thereby improving accuracy and reliability in entity extraction. Additionally, this chapter investigates computational approaches for extracting, disambiguating, and geolocating toponyms within Holocaust oral narratives. Toponym identification and resolution play vital roles in this context, as place names may be referenced in multiple languages (German, Polish, Yiddish, and Hebrew), may refer to locations that existed only during specific periods, or may have undergone significant name changes since World War II. The methods discussed here aim to not only identify these geographical references but also accurately map them to their historical coordinates and contexts.

4.1 Domain-specific Pre-Trained Language Models for Holocaust Oral Narratives

Encoder-only pre-trained language models (PLMs) are primarily designed for discriminative tasks, where the objective is to classify or extract information from text rather than generate it. These models learn contextual embeddings that capture rich bidirectional semantic dependencies between tokens. Unlike unidirectional models, which only consider past or future tokens, encoder-only PLMs leverage bidirectional attention mechanisms (via the Transformer encoder) to process the entire sequence simultaneously. This bidirectionality enables the model to make context-sensitive representations that change the meaning of a word or phrase based on the context around it. These representations excel in tasks requiring semantic understanding, such as Named Entity Recognition (NER)

and sentiment analysis, among others. Furthermore, encoder-only PLMs perform better in question answering, where identifying the precise answer span requires nuanced comprehension of both the query and passage. Furthermore, encoder-only PLMs serve as feature extractors that transform raw text into high-dimensional embeddings that encode syntactic, semantic, and relational information. These embeddings are able to be fine-tuned and adapted through lightweight architectures to improve downstream task performance while reducing computational overhead. Their strength lies in providing robust semantic representations, making them foundational for a wide range of applications in natural language understanding.

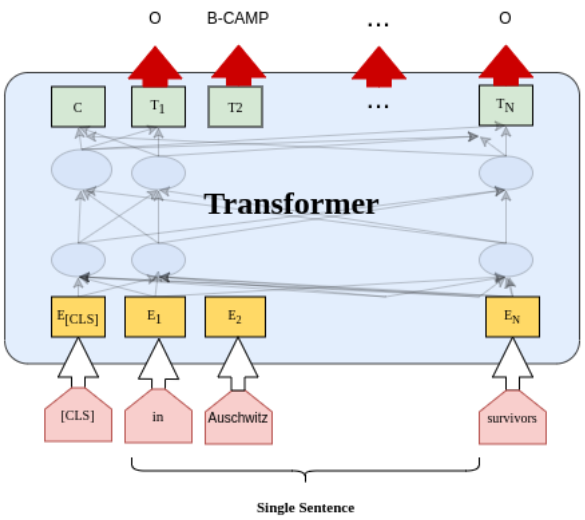

Figure 4.1: Encoder-Only Transformer Architecture for Holocaust Named Entity Recognition

This chapter discusses the proposed design, implementation, and evaluation of a domain-adapted NER framework for processing Holocaust oral narratives. The design phase focuses on utilising existing pre-trained language models and adapting the further pre-training and fine-tuning for the Holocaust testimony domain. Evaluation examines the model's performance and limitations compared to generic NER systems.

4.1.1 Corpus Preparation and Domain Adaptation

Aligning with our goal of developing domain-specific NER models for information extraction from oral narratives, the first step was to collect a substantial sample of English-language Holocaust testimonies. This corpus is used for both further pre-training and fine-tuning of language models. To ensure breadth and institutional credibility, we collected data from different major Holocaust documentation centres: the Wiener Holocaust Library, United Kingdom; the United States Holocaust Memorial Museum, USA; the Fortunoff Video Archive for Holocaust Testimonies at Yale University, USA; and the Centropa Archive funded by the European Union. The following criteria were used to choose Holocaust testimonials for our study:

- Language: Testimonial transcripts should be documented in the English language. While this constraint limits the linguistic scope of our current study, it establishes a controlled environment for our information extraction pipeline.
- Accessibility: Transcripts of testimonies must be publicly accessible in digital format to facilitate computational analysis.
- Completeness: Testimonies must present coherent and complete contextual information from the survivor's pre-war life through liberation or the immediate post-war period.

As discussed in section 03, of all the collected transcripts, we used 1,500 testimonies for further pre-training (MLM task), and the rest were used for the fine-tuning process. A set of 200 testimonies (human-annotated gold test set) was used for the evaluation purpose.

4.1.2 Model Architecture and Training Pipeline

Initial experiments utilising NLP tools demonstrated notable limitations in processing Holocaust testimony transcripts. Standard spaCy models, primarily trained on news and web texts, encountered difficulties in accurately recognising and

categorising domain-specific entities. Similarly, general-purpose deep learning models were unable to effectively incorporate the historical context inherent in survivor narratives. The poor performance of these traditional methods underscored the necessity for a domain-specific approach capable of addressing the linguistic and historical nuances inherent in Holocaust testimony. As a result of that, a two-phase approach is followed based on transformer architecture in this study for domain-specific entity recognition. First, a domain-adaptive further pre-trained language model was developed to teach the linguistic context of Holocaust testimonies. Subsequently, the resulting model was fine-tuned on a labelled NER dataset to perform the final token classification task.

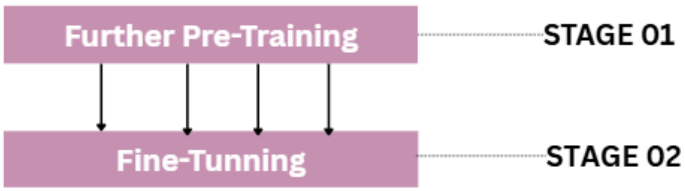

Figure 4.2: Two-Stage Domain Adaptation Training Process

4.1.2.1 Further-Pretraining Language Models for Domain-specific Oral Narratives

In the process of further pretraining, we have leveraged masked language modelling and Permutation Language Modelling (PeLM) learning objectives for the further-pretraining process. In MLM, a portion of words has been masked in the input text, and the model is trained to predict these masked words using the surrounding words. Meanwhile, in PeLM, a model predicts words by considering all possible ways a sentence can be ordered, helping it understand context more naturally and thoroughly.

For BERT, we maintained a masking probability of 15%, a batch size of 32, and a maximum sequence length of 512 tokens. Both models underwent further pre-

training for three epochs using the AdamW optimiser (learning rate: 2e-5, warmup steps: 10% of total steps). Training was executed on a single NVIDIA A100 GPU (40GB VRAM) with mixed-precision (FP16) training. This domain-adaptive further pre-training enables the model to learn the linguistic and contextual features in Holocaust testimonies, reducing domain mismatch errors observed in generic models. While perplexity provides an intrinsic measure of a model's fluency and familiarity with the domain language, the final goal is to achieve the best performance on downstream tasks. Therefore, to extrinsically evaluate the effectiveness of our further pre-trained models, we have assessed performance on the downstream task in domain-specific NER.

4.1.2.2 Fine-Tuning Language Models for Domain-specific Named Entity Recognition

A fundamental prerequisite for this fine-tuning process is a high-quality, domain-specific labelled dataset. Given the significant time and cost associated with manually annotating testimonies, we adopted a semi-automatic annotation approach, detailed in Section 3.3.2. We believe this approach balances annotation quality with practical resource limitations while maintaining the domain-specific accuracy. Further, it helps to enhance the precision and linguistically inherited features within Holocaust testimonies. In order to provide additional domain-specific knowledge, we integrated given external lexical resources into our semi-automatic annotation pipeline:

- Naval Vessels Database: A collection of warship names, including maritime-related entities in testimonies involving sea transport, naval operations, or maritime escape routes compiled from the World War II Database [1].

- Concentration Camp Registry: An authoritative catalogue of concentration camps, subcamps, ghettos, and transit facilities, providing standardised

[1] `https://ww2db.com/`

terminology and variant spellings of these locations as they appear in survivor testimonies. These lexical resources have been published and are maintained by the European Holocaust Research Infrastructure (EHRI)[2], ensuring their scholarly credibility and regular updates.

We hypothesise that integrating domain-specific lexicons will significantly improve the recall and precision of our semi-automatic annotation process. These lexicons capture key entities that are often overlooked by pattern-matching algorithms alone, resulting in a more comprehensive and accurate training dataset. The Inside-Outside-Beginning (IOB) format (Ramshaw and Marcus, 1999) was followed for this semi-automatic labelling approach. This format is essential for accurately representing multi-word entities (e.g., Lieutenant Joseph Goebbels-PER) by using special tags to indicate if a token is at the Beginning, Inside, or Outside of an entity.

For the fine-tuning process, various state-of-the-art transformer-based models were employed, including BERT (Devlin et al., 2019), RoBERTa, XLM-Roberta and XLNet. RoBERTa (Robustly Optimised BERT Pretraining Approach) is an advanced transformer-based language model that builds upon the BERT architecture(Zhuang et al., 2021). RoBERTa used dynamic masking, which randomly masks different subsets of tokens at each training epoch, allowing it to learn more effectively from diverse and complex data. Furthermore, XLM-Roberta (XLM-R)(Conneau, Khandelwal, et al., 2019), XLNet (Z. Yang et al., 2019) were designed to work with multiple languages. Additionally, another BERT-based language model (hmBERT), pre-trained on the historical domain corpora (Schweter et al., 2022), was used for this study. The multilingual BERT (mBERT) model was employed for the designed experiment, which performs well with multilingual settings.

[2] `https://portal.ehri-project.eu/`

4.1.3 Evaluation and Comparative Analysis

The performance of the domain-specific models was evaluated through a structured, two-stage assessment designed to measure their efficacy at key developmental and deployment phases. First, *intrinsic evaluation* was conducted to measure the capabilities of further pretrained models on the domain-specific corpus. Thereafter, *extrinsic evaluation* was followed by evaluating fine-tuned models for the NER task, which serves as the application for information extraction from Holocaust testimonies. The following comparative analysis is designed to validate the effectiveness of domain-adaptation strategies and identify the most suitable architecture for this historical domain.

4.1.3.1 Evaluation of Further Pre-trained Language Models

After further pre-training the PLMs on the Holocaust testimony corpus, their performance was intrinsically evaluated using Perplexity (PPL). Perplexity calculates how well the model predicts the next word in a sequence, and the model with lower perplexity is able to provide better model performance.

For a test set $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, perplexity is defined as:

$$\text{Perplexity}(\mathbf{x}) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(x_i)\right)$$

where:

- $P(x_i)$ is the probability assigned to x_i by the model,
- N is the number of samples in the test set.

According to this study, the model which was further pre-trained using the MLM objective was able to gain better results than the other models. Table 4.1 refers to the perplexity score of the further pretrained models.

The computational cost of pre-training for LLM from scratch is significantly higher than that of further pre-training due to the extensive resources required to

Table 4.1: Perplexity of further pretrained models

Model Name	Training objective	Perplexity
HoloBERT	MLM	3.1259
HoloRoBERTa	MLM	3.8178
HoloXLNet	PLM	8.975

train on large corpora. However, further pre-training, which adapts a pre-trained model to domain-specific contexts, is comparatively less resource-intensive, as it leverages the existing knowledge encoded in the model.

4.1.3.2 Results and Evaluation of the Domain-specific NER

To evaluate the performance of the NER model, we conducted several experiments by fine-tuning different PLMs and compared them with our further pre-trained model. We used standard evaluation metrics, the F-measure, to report and compare our experimental results. The F-measure (F_1) is defined as the harmonic mean of precision (P) and recall (R), given by

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{4.1}$$

Table 4.2 refers to the performance of individual named entity tags. According to our observation, Holocaust testimonies often contain ambiguous place names which require additional information to properly recognise. These ambiguous place names (toponyms), which are proper nouns, could have significantly different meanings based on the context of the testimony. For example, a survivor might say, “I was then sent to the labour camp in Czestochowa, which was called Hasag. " And in another place, "Czestochowa was a very difficult and brutal place to be." According to the context, "Czestochowa" is a toponym for a specific LOCATION in Poland where a labour CAMP was located during the Holocaust. This ambiguous nature related to place names could make it harder to identify and analyse the locations where events described in testimonies took place. This is a critical challenge faced during this study when analysing the tags such as LOC, GHETTO, and GPE.

Table 4.2: Evaluation Results: F1-Scores at the Entity Level for NER

Tags	**HmBERT**	**MACBERT**	**BERT**	**XLNET**	**HoloBERT**
B-CAMP	0.3812	0.3543	0.8954	0.8781	**0.8980**
B-DATE	0.8915	0.8919	0.9371	**0.9381**	0.9363
B-EVENT	0.4390	0.7652	0.7652	**0.7717**	0.7565
B-GHETTO	0.4211	0.4375	**0.9048**	0.8489	0.8529
B-GPE	0.7876	0.7870	**0.9395**	0.9380	0.9377
B-LANGUAGE	0.7116	0.6764	**0.8841**	0.8625	0.8745
B-LAW	0.5301	0.5195	0.7283	0.7553	**0.8045**
B-LOC	0.2677	0.2615	0.6369	0.6237	**0.6458**
B-MILITARY	0.0000	0.0225	0.7579	0.7225	**0.7513**
B-ORG	0.5843	0.5820	0.8457	**0.8525**	0.8450
B-PERSON	0.5666	0.5594	0.9179	0.9193	**0.9198**
B-RIVER	0.4938	0.5679	0.6237	0.6066	**0.7174**
B-SHIP	0.0000	0.0000	0.4000	**0.5714**	0.3333
B-SPOUSAL	0.5263	0.5641	0.8966	0.9153	**0.9153**
B-STREET	0.0683	0.0361	0.9143	0.9019	**0.9145**
B-TIME	0.8798	0.8760	0.8954	**0.8981**	0.8941
I-CAMP	0.3173	0.2680	0.7895	**0.7900**	0.7778
I-DATE	0.9308	0.9320	0.9476	**0.9505**	0.9491
I-EVENT	0.5007	0.4965	0.7785	**0.8048**	0.7994
I-GPE	0.4099	0.4240	0.7293	**0.7387**	0.7080
I-LAW	0.5793	0.6100	0.7360	0.7957	**0.8109**
I-LOC	0.1948	0.2357	**0.5863**	0.5714	0.5707
I-MILITARY	0.0258	0.0144	0.6759	0.6642	**0.6931**
I-ORG	0.6416	0.6442	0.8320	**0.8410**	0.8310
I-PERSON	0.6435	0.6350	0.9124	0.9108	**0.9124**
I-RIVER	0.4471	0.4889	0.6374	0.6415	**0.7071**
I-SPOUSAL	0.5577	0.5739	0.8143	0.8212	**0.8188**
I-STREET	0.1429	0.1569	0.9444	0.8974	**0.9577**
I-TIME	0.9110	0.9076	0.9204	**0.9228**	0.9198

Therefore, the next section explores further the possible approaches which can be tried out to mitigate the toponym disambiguation.

4.2 Toponym Disambiguation in Holocaust Narratives

Toponym resolution in NLP is a challenging task which remains unaddressed (Z. Zhang, Laparra, and Bethard, 2024; X. Hu, Kersten, and Klan, 2025). The complexity of the task has been exacerbated by the fact that, over time, geographic locations have been referred to by different names in textual documents. However, in the context of the spoken text, a single name can refer to two distinct locations in historical documents. In Holocaust testimonies, survivors use "Auschwitz" interchangeably to describe the town where they were initially brought and the camp where they were imprisoned. In order to disambiguate the toponyms, previous studies used different databases which represent the coordinates of the geographical places (Gritta, Pilehvar, and Collier, 2020; Sá, Da Silva, and Macêdo, 2022). However, the specific coordinates of the locations relevant to the Holocaust were not included in the aforementioned databases because those locations were only mentioned within the Holocaust domain.

Compared to written texts, the unstructured nature of transcribed oral narratives introduces ambiguities that complicate the understanding of contextual meaning. In Holocaust testimonies, survivors explain what they have witnessed during the World War period, and frequently transcribed testimonies consist of sensitive and traumatised experiences of survivors in different geographical locations. Due to the traumatic nature, testimonies often lack consistency in naming conventions, highlighting the need for NER systems capable of resolving toponyms.

In this study we discuss the toponym resolution task by employing different prompt engineering techniques, such as retrieval-augmented generation and few-shot Chain-of-Thoughts (COT). While some research has been conducted on the

Example 01: Referring the same name for different contexts

We	were	taken	to	**Theresienstadt**	transit	camp	to	Majdanek
O	O	O	O	B-CAMP	O	O	O	B-CAMP

All	of	us	stayed	in	**Theresienstadt**	for	three	nights
O	O	O	O	O	B-GHETTO	O	O	O

Example 02: Different spelling referring to the same place example (**Auschwitz- Birkenau is a one camp**)

who	had	to	come	to	**Auschwitz**	in	1942	from	Slovakia
O	O	O	O	O	B-CAMP	O	B-DATE	O	B-GPE

those	unfit	for	further	experiments	were	sent	back	to	**Birkenau**	or	gassed
O	O	O	O	O	O	O	O	O	B-CAMP	O	O

Example 03: Symbols refer the geographical location

They	were	transported	to	**KZ**	Flossenbuerg	in	Bavaria
O	O	O	O	B-CAMP	I-CAMP	O	B-GPE

Figure 4.3: Representative Sentence Examples Extracted from Holocaust Testimonies

recognition of geographical features in the Lake District Corpus (Ezeani, Rayson, I. Gregory, et al., 2023; Ezeani, Rayson, and I. N. Gregory, 2023), no comparable effort has been directed toward the systematic identification and extraction of such entities within Holocaust testimony collections. However, (Ezeani, Rayson, I. Gregory, et al., 2023; Ezeani, Rayson, and I. N. Gregory, 2023) was unable to reproduce and not capable of assessing whether the toponym is a man-made location such as a bridge, building, or house, or a natural location such as a river, forest, or sea. Figure 4.3 provides real examples of the ambiguities of the NE in the testimonial contexts.

4.2.1 Corpus Preparation and Domain Adaptation

For the toponym resolution task, the same dataset was employed with the same format as explained in the NER task. The data maintains the BIO tagging scheme, where locations previously identified by the NER system (tagged as B-LOC and I-LOC) serve as input for the toponym resolution pipeline. Having a standardised format of data ensures experimental consistency and reflects a realistic NLP pipeline

where entity recognition precedes entity linking and disambiguation.

4.2.2 Model Architecture and Training Pipeline

The traditional deep learning models are limited in toponym disambiguation by their static knowledge and inability to reason over external context. Therefore, we hypothesise that LLMs are able to overcome these limitations through dynamic, knowledge-augmented reasoning via Retrieval-Augmented Generation (RAG) and carefully engineered prompts. To compare general-purpose LLMs' effectiveness in the geospatial domain, two different prompting strategies, few-shot (FCOT) and zero-shot (ZCOT), were used. When refining the model in ZCOT, we attempted to obtain responses straight from the LLM, whereas in FCOT, we used the labelled knowledge base as the retriever. GPT-4o and Llama 3.0 (Llama-3-70B-Instruct) were used as the models for the experiments of this study by setting up the temperature to 0 and a maximum token limit of 1500 for each output.

Approach 01: Zero-shot COT (ZCOT) prompting

Zero-shot prompting (ZCOT) is the prompt engineering technique where language models generate responses for tasks without any new examples or fine-tuning. In the prompt given in Figure 4.4, we have used the following information to enhance geospatial knowledge within the prompt during the inference process:

- LOC: Locations except countries or cities.
- GPE: Geographical locations such as countries or cities.
- CAMP: Concentration camps (Extermination, Transit, Labour)
- GHETTO: Ghettos, the Jewish quarters in cities.
- STREET: Pathways or roads.

However, after experimenting with the ZCOT method, we concluded that the prompt needed to be more advanced with more domain-specific information in order to be accurate in identifying geospatial entities such as GHETTO.

```
Zero-shot Chain-of-Thought Prompt for Holocaust Testimonies Corpus
Context: Consider the temporal scope from 1936 to 1944.
Task: Identify name entity tags for holocaust-specific tags.
The list of name entity tags is: {list_of_tags}.
Each tag is defined as follows: {tags_meaning}.
Instructions:
1. Identify the most suitable name entity tag for the word 'NAMEENTITY' in the
    ↪ GIVEN SENTENCE based on the criteria:
   Analyze the words preceding and following 'NAMEENTITY'.
   Understand the complete sentence context and identify specific factors related
       ↪  to the target word.
   The GIVEN SENTENCE: {sentence}.

2. Return only the GIVEN SENTENCE after assigning the identified tag to '
    ↪ NAMEENTITY'.
   Do not add any additional data or explanation.
Output Format:
"<Updated sentence with correctly identified name entity tags>"
```

Figure 4.4: Zero-Shot Chain-of-Thought Prompt for Named Entity Recognition on the Holocaust Testimony Corpus. The placeholders `{list_of_tags}`, `{tags_meaning}`, `{sentence}`, and `NAMEENTITY` are replaced with actual values during execution.

Approach 02: Retrieval Augmented Generation (RAG)

In the RAG approach, we shared the geospatial knowledge which was not available in the training set during the prompting process, which resulted in better and more accurate responses. The proposed design includes two phases: vector store generation and the retriever with response generation with the selected language models. In this study the following models were used.

- Vector store generation and embedding: The 'BGE small' model from Hugging Face was chosen as the embedding for the study, while Chroma DB was

employed to store the vectors related to the labelled geospatial data. To preserve contextual meaning during chunking, a recursive character text splitter from LangChain was incorporated to create the necessary data chunks with 2500 tokens, overlapping 50 tokens. These chunks are stored in the vector store once embedded using the embedding model.

- Retriever and prompting: Retrieval QA was utilised to build the retriever, with search_kwargs(k) set to '2' and the search_type set to 'similarity'. The similarity search uses cosine similarity to extract the vectors closest to the input sentence we want to tag. This approach allows us to feed data with a similarly labelled context to the model, enriching the response generation task with geospatial knowledge. The designed FCOT prompt is employed here, with minor adjustments to fit it into the process. We observe that the main flaw of this approach is that the retriever uses similarity score assessments to retrieve related data based on the sentence context. As a result, sample chunks without the sought word may be returned. To address the aforementioned issue, we modified the key-word based retriever in the next method to use the knowledge base.

We observe that the main flaw of this approach is that the retriever uses similarity score assessments to retrieve related data based on the sentence context. As a result, sample chunks without the sought word may be returned. In order to address the limitation, in the next approach we improved the retriever to utilise the knowledge base when generating the results.

Approach 03: Few-shot COT prompting with Knowledge Base

In this approach, the pre-labelled knowledge base is incorporated into the inference process by FCOT. In order to obtain the few-shot prompts required for inference, we stored labelled geographical entities in a knowledge graph. To generate responses, the stored knowledge is utilised. The tree structure of the knowledge graph is designed with the place names as the root node and geospatial entities as the first-level parent nodes. Leaf nodes are designed as the list structure containing

sample instances of labelled datasets. Figure 4.5 demonstrates the structural view of the knowledge base.

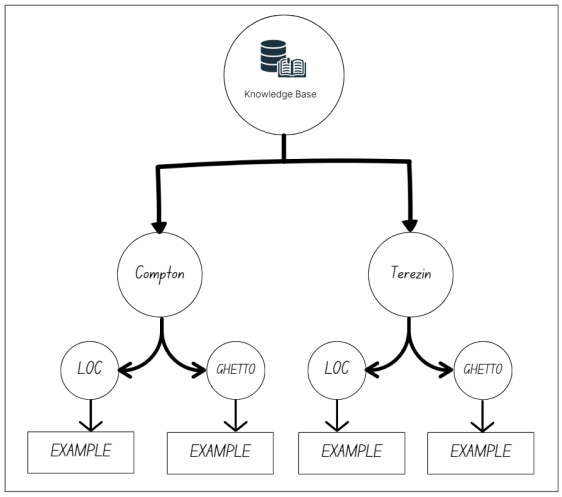

Figure 4.5: Knowledge Base Structure for Toponym Disambiguation

The above structural definition of knowledge graph has effectively improved the retrieval time of example phrases required for few-shot learning, which can be performed effectively. Further, the presence of the target word in the retrieved sentences is considered mandatory for efficient labelling in the few-shot approach. At most five instances for each entity are retrieved from KB. If the word is absent, the prompt will function in a zero-shot manner. The detailed workflow of the approach is presented in Figure 4.6.

4.2.3 Evaluation and Comparative Analysis

Table 4.3 presents the evaluation of the baseline model with ZCOT prompting, the RAG approach and the FCOT approach with an additional knowledge base. Compared to proprietary and open-source LLMs, the GPT-4o proprietary language model outperforms the Llama open-source language model with the same parameter setting. Based on our experiments, the pure ZCOT approach underperforms at recognising domain-specific entities. As the second approach, we combined the RAG pipeline, which targets sentence-level retrievers using the cosine distance. Compared

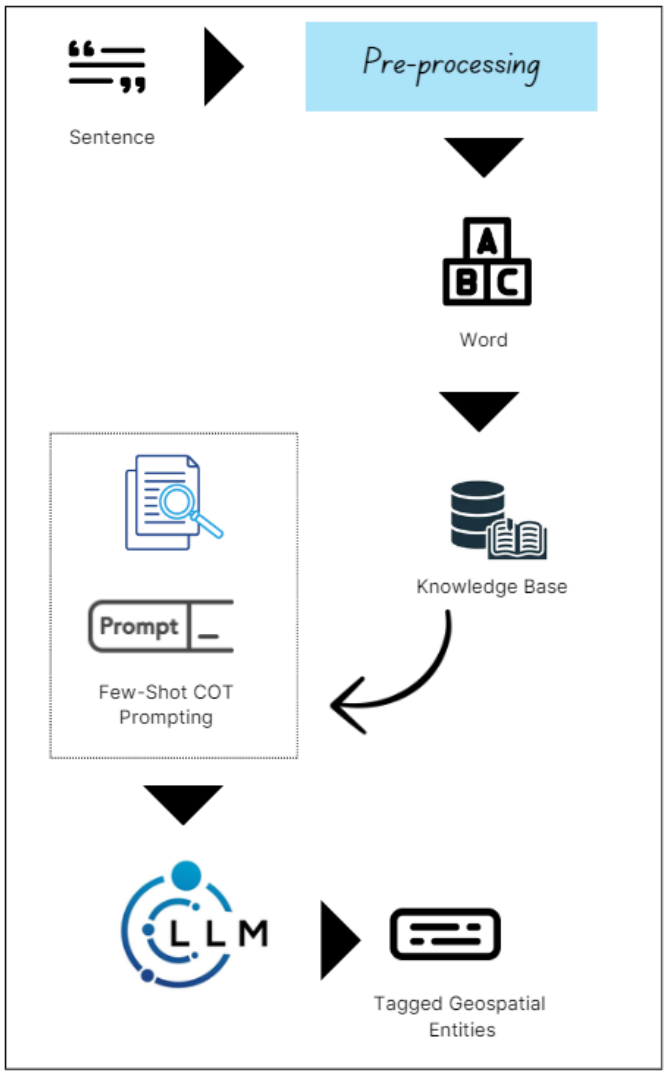

Figure 4.6: Data Flow of the Few-Shot Chain-of-Thought NER Pipeline

to the baseline approach, recognising the GHETTO tag shows a 0.15 improvement in F1 score, while other entities show a slight improvement in the tagging. This approach has shown that proper retrieval would improve the performance of the tagging process.

The lack of geospatial knowledge is then addressed using a model using the FCOT approach. The word-orientated retriever, which uses a structured knowledge base, is incorporated to extract the most appropriate result. The FCOT approach has shown significant improvements for the GHETTO tag, with an increase of 0.19%

Table 4.3: Performance comparison between prompt engineering techniques. (GPT-4o)

Entities	Spacy Rule based model			Zero-shot GPT 4o			RAG with GPT 4o			Few-shot COT Prompting		
	Precision	Recall	F1score	Precision	Recall	F1score	Precision	Recall	F1score	Precision	Recall	F1score
LOC	1.0	0.10	0.18	0.57	0.74	0.64	0.54	0.81	0.64	0.63	0.84	**0.72**
GPE	0.64	0.83	0.72	0.77	0.85	0.81	0.83	0.77	0.80	0.89	0.82	**0.85**
CAMP	0.77	0.51	0.62	0.95	0.74	**0.83**	0.90	0.76	0.82	0.88	0.79	**0.83**
GHETTO	0.00	0.00	0.00	0.62	0.31	0.41	0.59	0.53	0.56	0.61	0.59	**0.60**
STREET	0.67	0.51	0.58	0.94	0.84	0.88	0.97	0.94	**0.95**	0.91	0.91	0.91

Table 4.4: Performance comparison between prompt engineering techniques. (Llama)

Entities	Spacy Rule based model			Zero-shot Llama			RAG with Llama			Few-shot COT Prompting		
	Precision	Recall	F1score	Precision	Recall	F1score	Precision	Recall	F1score			
LOC	1.0	0.10	0.18	0.54	0.49	0.51	0.40	0.59	0.47	0.44	0.56	0.49
GPE	0.64	0.83	0.72	0.78	0.78	0.78	0.76	0.79	0.77	0.81	0.75	0.78
CAMP	0.77	0.51	0.62	0.89	0.77	0.83	0.81	0.74	0.77	0.84	0.73	0.78
GHETTO	0.00	0.00	0.00	0.65	0.50	0.57	0.43	0.69	0.53	0.45	0.47	0.46
STREET	0.67	0.51	0.58	0.73	0.91	0.81	0.74	0.92	0.82	0.86	0.90	0.88

in the F1 score, and for the LOC category, with an improvement of 0.08% in the F1 score. Our findings demonstrate that well-crafted prompts, along with a knowledge-sharing approach, can assist the general-purpose LLM to extract domain-specific complex tasks such as toponym resolution. Table 4.3 demonstrates that effective prompt engineering alone can successfully resolve more highly ambiguous named entities than rule-based approaches in the table, yielding substantial performance gains without any fine-tuning.

4.3 Domain-specific challenges of NER respect Holocaust Testimonies

Holocaust testimonies differ substantially from the textual domains on which most NER models are trained, exposing methodological, linguistic, and ethical challenges. One major challenge arises from the linguistic characteristics of Holocaust testimonies. Many Holocaust testimonies originate from oral narratives

that have been transcribed later, leading to fragmented syntax, repetitions, interruptions, and incomplete sentences. Survivors use implicit references, such as pronouns or vague expressions like "there" or "the camp", assuming shared contextual knowledge with the interviewer or the reader of the testimonies.

Furthermore, testimonies often contain emotional digressions and non-linear storytelling, which further complicate automatic entity recognition. More information is discussed in the section 2. Historical specificity also poses substantial challenges for NER in this domain. Holocaust testimonies include references to places whose names have changed over time, no longer exist, or are geographic entities that have shifted borders and boundaries. Survivors frequently use colloquial, outdated, or locally used place names which are absent in modern world maps. Similarly, organisations and institutions referenced in testimonies, such as Nazi administrative units or concentration camp structures, couldn't be recognised directly through the standard NER categories. Another important issue concerns the mismatch between standard NER taxonomies and the entity types present in Holocaust testimonies. While traditional NER focuses on categories such as person, location, and organisation, Holocaust narratives frequently reference domain-specific entities such as ghettos, concentration and extermination camps, military vessels, and events. These entities carry historical meanings that cannot be adequately captured by coarse-grained entity labelling. Additionally, nested and overlapping entities, such as compound camp names, further challenge automatic sequence-labelling approaches in NER.

Data scarcity and annotation complexity further limit progress in this area. Existing datasets for the Holocaust testimony domain are often small, institutionally restricted, or annotated according to inconsistent guidelines. Moreover, annotation in this domain requires not only linguistic competence but also historical knowledge and ethical sensitivity. Because incorrect entity labelling risks distorting survivor narratives and historical facts, increasing the cost and difficulty of creating reliable training data.

Finally, Holocaust testimonies exhibit significant temporal and referential complexity. In narratives survivors frequently shift between different time periods, blending past experiences with retrospective reflections. Personal names may change over time due to migration or marriage, or individuals may be referred to by nicknames or roles rather than their proper names. Standard NER systems lack mechanisms to account for such temporal shifts and evolving entity references, which limits their ability to maintain consistent entity recognition across a testimony. In conclusion, Holocaust testimonies constitute a uniquely challenging domain for NER due to their oral language, multilingual nature, historical specificity, domain-specific entity types, and ethical constraints. The above highlighted challenges demonstrate the inadequacy of general-purpose NER systems for this area and underscore the need for domain-adapted models with extended annotation schemes. Addressing these challenges is essential for enabling reliable information extraction from Holocaust testimonies and for supporting broader research in digital humanities and Holocaust studies.

4.4 Chapter Summary

This chapter provided an in-depth discussion on adapting a domain-specific NER system for historically significant documents, particularly focusing on Holocaust survivor testimonies. Different technical adaptations were explored within the paradigm of language modelling, including the further pretraining and fine-tuning of PLMs such as BERT on Holocaust-specific corpora. In addition, supplementary tools and strategies were proposed to address limitations in traditional NER systems, such as integrating toponym resolution techniques to disambiguate historically referenced place names. By incorporating domain adaptation techniques and leveraging the power of LLMs, the chapter demonstrated how advanced NER systems can support the accurate identification of named entities from the testimonies. This work contributes to improved entity recognition performance and

facilitates deeper historical analysis, digital preservation, and access to survivor narratives in meaningful and context-aware ways. To the best of our knowledge, this is the first study seeking to accurately identify and disambiguate toponyms which denote the geographic representations of Geopolitical Entity (GPE), Location (LOC), Concentration Camp (CAMP), Ghetto (GHETTO) and Street (STREET) mentioned in transcribed text and related to specific historical events such as the Holocaust.

Chapter 5

Modelling Relationships in Historical Narratives

I feel so strongly that I survived to be the voice of the thousands and thousands of children who died during the Holocaust.

Ellen Litman(Survivor)

The previous chapter presented the current work in applying NER techniques to domain-specific oral testimonies, with a focus on identifying entities, such as people, locations, and events embedded in survivor narratives. This research shows that automated information extraction methods are crucial for organising unstructured testimonial data. However, entity recognition alone is insufficient to capture the semantic connections and interactions that exist between them. As a result, the connections among people, places, and events that shaped survivors' experiences remain underexplored. To address this gap, this chapter introduces the next stage of analysis: relationship extraction. This method goes beyond simple entity recognition by mapping the semantic relationships that connect entities, tracing pathways such as geographical movements or kinship ties. By mapping such connections, relationship extraction helps to analyse the Holocaust testimonies more precisely, revealing patterns of human memory.

5.1 Domain-specific Relationship Extraction

Relationship Extraction (RE) is a fundamental component in the field of NLP, serving to uncover meaningful connections and associations between entities within textual data. RE has been broadly studied in the last few decades, with many datasets published across different domains, and some studies classified into three data sources: (1) news and web, (2) scientific publications and (3) Wikipedia (Bassignana and Plank, 2022). However, the application of RE within the digital humanities has received less scholarly attention than in other areas. In the context of Holocaust testimonies, RE enables the discovery of historical connections, interpersonal relationships, and contextual associations hidden within archives of survivor narratives. As early as 2018, scholars identified significant inadequacies in the digital infrastructure for Holocaust studies (De Leeuw et al., 2018), challenges that persist to this day (Digital Holocaust Memory Project, 2023). Thus, RE creates an urgent need for computational approaches to effectively process and extract knowledge from these testimonies, where it represents not merely a technical challenge but a moral imperative to preserve and make accessible the experiences documented in these testimonies.

Some preliminary work has been conducted using rule-based computational approaches on multiple Holocaust victim reports to extract biographical information (Sagi et al., 2016).This study demonstrated the feasibility of applying computational methods to Holocaust-related documents. Yet it remained with limited scope, as it did not extend to survivor testimonies, which are known to contain richer relational information. This limitation is significant because rule-based systems often lack the flexibility and generalisation capability needed to handle the linguistic diversity and complexity of personal testimonies. The field of RE has evolved considerably, developing various approaches to identify and classify relationships between entities. Formally, given a domain-specific language text x, the task is to predict a set of relation triplets:

$$T = \{(e_1, r, e_2) \mid e_1, e_2 \in E,\ r \in R\} \tag{5.1}$$

Where e_1 and e_2 denote the head and tail entities, respectively, E is the set of all entities identified in x, and r is a predefined relation type such that $r \in R$, the set of all possible relation types. The entities e_1 and e_2 may be words, phrases, or syntactic units, which r capture the semantic relationship between them.

According to our observations, two primary approaches could be utilised for the Holocaust testimonies and other narrative documents. The first approach focuses on determining the existence of relationships between entities (Perera, Dehmer, and Emmert-Streib, 2020). This task focuses on identifying meaningful semantic relationships that exist between two entities in proximity within the text or whether they are merely co-mentioned without a specific connection, as a binary classification task. This approach could identify cases where two people are mentioned in the same context because they shared a meaningful relationship (such as being family members, fellow prisoners, or members of the same resistance group). Extracting predicate verbs as relationship types is the second approach to performing RE (Etzioni et al., 2011). Unlike domain-specific RE tasks that work with a closed set of predefined relationship categories, this open-ended approach treats any predicate verb that appears in a sentence and indicates a relationship between entities as a valid relationship type. This approach could capture the full range of relationships described in testimonies without being constrained by predefined RE categories. However, the method requires subsequent normalisation and classification to identify diverse relationship types.

Triplet Extraction

Triplet extraction and relationship extraction are interconnected concepts, where triplet extraction represents a specific structural output format of the relationship extraction task in NLP and information extraction (D. Li et al., 2025). In triplet extraction, as described in Figure 5.1, each extracted triplet consists of three components that together represent a complete semantic relationship. The subject,

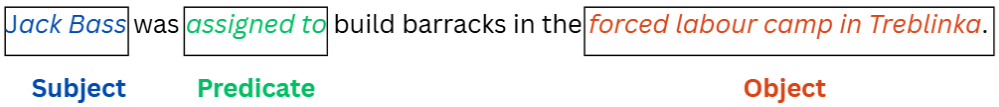

Figure 5.1: Example of a Relational Triplet Extracted from Holocaust Testimony ((Boder, 1946))

or head entity, represents the primary entity that acts, contains a property, or participates in a relationship. The predicate describes the connection between the subject and object, representing actions, states, properties, or associations. The object or tail entity represents the entity that receives the action, is affected by the relation, or completes the predicate, encompassing the same types of entities, including temporal expressions, numerical values, and abstract concepts.

In the context of Holocaust testimonies, subjects could include individuals such as survivors or perpetrators, organisations such as resistance groups or Nazi institutions, locations including ghettos and camps, or abstract entities representing events, policies, or conditions. Predicates could be the verb-based actions and events, nominal relations that express properties and attributes, or prepositional relations that indicate spatial, temporal, or associative connections related to the Holocaust.

Table 5.1: Primary categories of relations

Relationship Category	**Relationship**
Biographical	born, die, learn, live, locate, married
Career	work, employ, travel, return
Holocaust Events	forced, transport, evacuate, arrest, deport, kill

Understanding the distinction between open information extraction and traditional relation extraction is important for the field of digital humanities, particularly in the context of Holocaust testimonies. Traditional relationship extraction approaches require predefined sets of relations and annotated training data, which limits the prediction of domain-specific relationships, specifically when survivors

describe unique experiences, contextual connections, and historical relationships. Open relationship extraction overcomes these limitations by discovering relations without predefined schemas (X. Zhao et al., 2024; Pai et al., 2024) and extracts any relationship described in the text, regardless of whether it matches expected patterns (Siciliani et al., 2024). This approach is particularly important for testimony processing, as it can capture unique historical relationships, personal experiences, and contextual information that would be impossible to anticipate in a closed schema.

Employing computational approaches to open relationship extraction has evolved through several paradigms. Rule-based models utilise syntactic patterns to identify and extract relations, focusing on verb-mediated connections between noun phrases or extending to nominal and adjective-based relations. Dependency parse-based approaches use grammatical structure, using clause identification and dependency relations to extract predicates and their arguments with greater linguistic sophistication. Recent advances in Open Information Extraction (OpenIE) increasingly rely on end-to-end neural architectures and transformer-based models to improve extraction quality and scalability (S. Zhou et al., 2022). These neural approaches offer superior handling of complex sentences, better detection of implicit relations, and improved performance on diverse linguistic constructions (Pai et al., 2024).

The relation set R is restricted to three categories of relationships:

$$R_{\text{Holocaust Testimony}} = \{r_{\text{bio}}, r_{\text{geo}}, r_{\text{eve}}\}$$

where:

- r_{bio} represents biographical relations
- r_{geo} represents geographical relations
- r_{temp} represents event-specific relations

The task can be formalised as learning a mapping:

$$f\colon x \mapsto T_{\text{Holocaust}} \subseteq \mathscr{E} \times R_{\text{Holocaust}} \times \mathscr{E}$$

Which identifies only those triplets relevant to the Holocaust knowledge domain.

The technical implementation for mapping natural language to relation triplets involves several sequential computational steps. First, entity recognition and classification identify all relevant entities within the text. Next, to understand how these entities interact, the grammatical and semantic structure of sentences is analysed through dependency parsing or semantic role labelling. A critical distinction in this phase is the approach to defining relationships: in traditional, closed-domain extraction, relationships are classified into a fixed set of predefined categories, whereas in Open Information Extraction (OpenIE), relationships are extracted as open-ended linguistic expressions, which are then normalised. Finally, triplet construction assembles entities and their relationships into structured subject-predicate-object triples. This final step must also handle complex cases, such as generating multiple triplets from a single clause or resolving relationships that span across multiple sentences.

Figure 5.2 illustrates the proposed information extraction pipeline for relationship extraction. The proposed knowledge graph consists of four components: 1) Data processing, 2) Coreference resolution, 3) Triplet extraction 4) Visualisation. After the collection of Holocaust testimonies, the coreference resolution component identifies chains of entities and pronouns that refer to the same entity. The triplet extraction component extracts relation triplets from the text using open information extraction techniques, and lastly, extracted relationships are visualised in our findings on a graph database. The details of each component are presented below.

5.1.1 Corpus Preparation and Data processing

The development of a relationship extraction model for Holocaust testimonies requires a careful corpus preparation process to address the unique linguistic, historical, and semantic characteristics. Processing Holocaust testimonies presents distinct challenges, including multilingual and transliterated terminology, as well

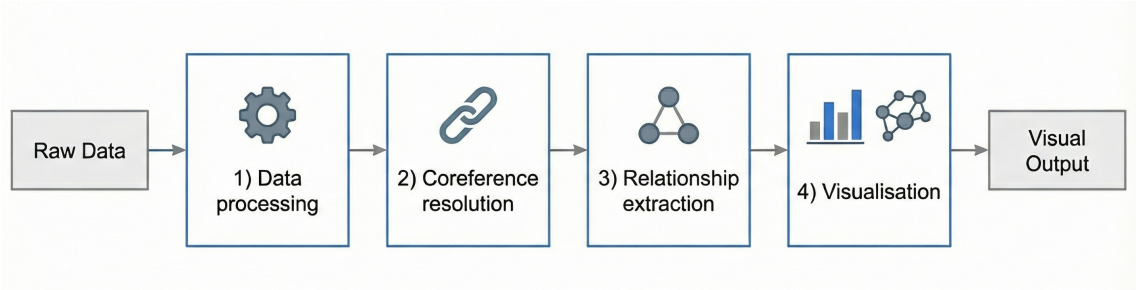

Figure 5.2: Pipeline for Relationship Extraction from Holocaust Testimonies

as complex, domain-specific historical relationships. The annotation process of Holocaust testimony corpora for relationship extraction presents unique challenges that distinguish it from more commonly studied text types. Unlike general-purpose text, where relationships are expressed in straightforward declarative sentences, Holocaust testimonies include complex narrative structures with temporal shifts, memories, and emotionally charged sentences that complicate the identification of relationships. As explained in the section 3 human annotated testimonies were used as the golden dataset.

5.1.1.1 Co-reference resolution

Coreference resolution involves identifying and clustering all linguistic expressions in a discourse that refer to the same real-world entity (Ruicheng Liu et al., 2023). In computational linguistics, coreference resolution addresses the challenge in natural language, which is inherently referential, as speakers and writers use various linguistic variations such as pronouns, definite descriptions, proper names, and nominal phrases to refer to entities without constant repetition. For Holocaust testimonies, coreference resolution is significant as survivors narrate their personal experiences and what they have witnessed involving multiple individuals, locations, and events across extended time periods, often using pronouns and varied descriptions to refer to the same people and places. In theory, coreference resolution encompasses various types of referential relationships. Pronominal

anaphora represents the most common type, where pronouns refer back to previously mentioned entities in the discourse. Nominal coreference explains when different noun phrases are used to describe the same entity, for example, when a person is first introduced by their full name and later referred to by their profession or a descriptive phrase. Further, proper name variations involve different forms or spellings of the same name, which is particularly relevant in multilingual contexts where names may be transliterated differently.

A coreference cluster represents the complete set of all mentions that refer to a single entity throughout a document. In knowledge graph construction, each coreference cluster maps to a single node, ensuring that all information about an entity is consolidated rather than fragmented across multiple disconnected nodes. These clusters could be extending the domain of narrative text, including dozens of mentions through pronouns, descriptive phrases, and name variations. However, the challenge is to accurately identify which mentions belong together when multiple entities of the same gender or type are discussed in close proximity when the discourse shifts between different time periods and contexts. Coreference resolution has undergone significant evolution over the past few decades, leading to state-of-the-art neural network-based architectures (Ruicheng Liu et al., 2023). These models learn different patterns directly from text, leveraging contextualised embeddings rather than traditional models. Their ability to handle long-range dependencies, learn complex patterns, and support multilingual processing makes them effective for the different task of resolving references in testimony narratives.

In Holocaust testimonies, the emotional weight of survivors' personal experiences often leads to fragmented discourse patterns, including incomplete sentences and repetitions. This fragmentation makes coreference resolution particularly significant, as it poses a distinct challenge to standard parsing algorithms. In testimonies, it is common for multiple entities with similar references, such as several family members or individuals, to be referred to by the same pronouns or relational terms when they are mentioned in close textual proximity. Multilingual and transliterated

names introduce complexity, as Hebrew, Yiddish, and other languages may have multiple romanised spellings, and names may appear in various forms across different parts of a testimony. Moreover, temporal features in testimonies, where speakers jump between different time periods in their lives, require maintaining entity continuity while tracking changing contexts. Moreover, the impact of coreference resolution on knowledge graph construction is direct because a knowledge graph may contain multiple disconnected nodes representing the same entity, leading to fragmented relationship networks and incomplete entity profiles. Information about a single person might be scattered across several nodes, making it impossible to retrieve a complete picture of that person's experiences and connections. Accurate coreference resolution allows all mentions to be connected to the nodes, ensuring relationships are properly attributed and consist of queryable representations of the testimonies. This consolidation is essential for enabling meaningful analysis and allowing researchers to trace connections between individuals, events, and locations across multiple testimony accounts.

As discussed above, Holocaust testimonies contain numerous types of pronouns that require resolution. Personal pronouns such as `he, she, and they` are the most obvious, appearing when the witness discusses what a family member or historical figure did. Possessive pronouns like `his, her`, and `their` are equally important because they establish ownership or relationship, as in `his wife` or `their children`. Demonstrative pronouns like `this` and `that` appear when referring to recently mentioned locations or events, such as `This city was destroyed.` Family-relation pronouns are important in testimony contexts where phrases such as `my mother, my grandmother, my father's brother`, and `his daughter` frequently appear and must be resolved to actual person names. Relative pronouns like `who, which`, and `that` connect subordinate clauses to specific entities, as in `My uncle who was a doctor`. The narrative and unstructured structure of the testimonies complicates the understanding of these complex scenarios.

Three primary approaches were tried out to address pronoun resolution issues,

Algorithm 1 Algorithm used for Coreference resolution (Pronoun resolution)

Require: $\mathcal{D}$: directory of testimony files T, each with relationships R_T

$\mathcal{D}_{\text{cleaned}} \leftarrow \emptyset$
for each testimony T in $\mathcal{D}$ **do**
  $\mathcal{FM} \leftarrow$ LLM_FamilyMapping(S)
  for each relationship rel in R_T **do**
    if IsKinshipTerm(rel.subject) **then**
      rel.subject $\leftarrow \mathcal{FM}$[GetKinshipRole(rel.subject)]
    end if
  end for
  Save T' to $\mathcal{D}_{\text{cleaned}}$
end for

- LLM_FamilyMapping(S): Queries DeepSeek API to obtain family mapping $\mathcal{FM}$
- IsKinshipTerm(*entity*): Detects kinship keywords
- GetKinshipRole(*entity*): Maps entity text to kinship role (mother, father, etc.)

each with different trade-offs between accuracy, cost, and scalability. As the first approach, rule-based approaches were followed by the open-source NLP libraries designed for coreference resolution, such as AllenNLP's coreference resolution models and spaCy's neural coreference resolution. In a rule-based approach, linguistic rules were defined to resolve common pronoun patterns based on Part-of-Speech (POS) tags. For example, a rule that says, When you encounter 'my', replace it with the actual name of that family member, or another rule could state. Within the same sentence or paragraph, if 'he' appears after mentioning a male person's name, the 'he' refers to that male person.

As the second approach, machine learning-based NLP models (AllenNLP's coreference resolution models and spaCy's neural coreference) were employed, which were trained on large, annotated datasets to automatically identify which pronouns

and noun phrases refer to the same entity. The advantage of this approach is that it's fully automated and can process large volumes of text quickly. The disadvantage is that accuracy varies depending on domain-specific language in testimonies (models trained on news articles might not perform as well on historical testimony language). Furthermore, our experiments have revealed that the highly volatile nature of the testimonies, their unstructured grammar, and the fact that each one differs from the other testimonies complicate the process of pronoun resolution. Therefore, the only approach that yielded good results was the LLM-based approach (the prompt is illustrated in Figure 5.3). This involves providing the entire testimony to the LLM and instructing it to identify and resolve all relevant pronouns, which it accomplishes through its deep understanding of contextual language.

Following the completion of the coreference resolution, we conducted experiments with multiple relationship extraction techniques to discover which one performed the best for the Holocaust testimonies.

5.1.2 Model Architectures and Training Pipeline

In this study, as illastrated in Figure 5.4, traditional rule-based methods were utilised with part-of-speech tagging and handcrafted rules for identifying subject-predicate-object patterns. According to this, relationships were identified as linguistic patterns through syntactic structures, where subjects of phrases appeared as noun phrases in specific grammatical positions, predicates were expressed through verbs or verb phrases, and objects appeared as noun phrases in complement positions. Handcrafted rules were constructed based on part-of-speech sequences and dependency patterns that frequently indicate relationship expressions. However, this approach had significant flaws when applied to testimonies because of the unstructured, fragmented, long-range dependencies and non-canonical word orders resulting from emotional expression, which prevented reliable pattern matching. The diversity of relationship types included in testimonies, encompassing not only factual historical relationships but also emotional and interpretive connections, could not

```
You are an expert in extracting family relationships from oral history
    ↪ testimonies.
From the narrator's ("I","my") point of view, identify the real names of these
    ↪ family members.

Return ONLY a clean JSON object with exactly these keys when found (do NOT
    ↪ include nulls or missing ones):

{
  "mother":"Full Name",
  "father":"Full Name",
  "grandmother":"Full Name",// prefer the most frequently mentioned one
  "grandfather":"Full Name",
  "sister":"Full Name",
  "brother":"Full Name",
  "daughter":"Full Name",
  "son":"Full Name"

Use the most common or full version of the name as it appears in the text.
If multiple names exist for one role (e.g. maiden + married), use the primary one
    ↪  used in the story.
Only include a key if you're 100% sure.

Furthermore replace the 'Here', 'There' by the proper location or desirable
    ↪ entity.
"""

Holocaust testimony: ```{para}```
```

Figure 5.3: Zero-Shot Prompt for Coreference Resolution

be captured through predefined syntactic rules.

The second approach attempted to use the AllenNLP framework using argument mining techniques to identify and extract relationship arguments from testimony

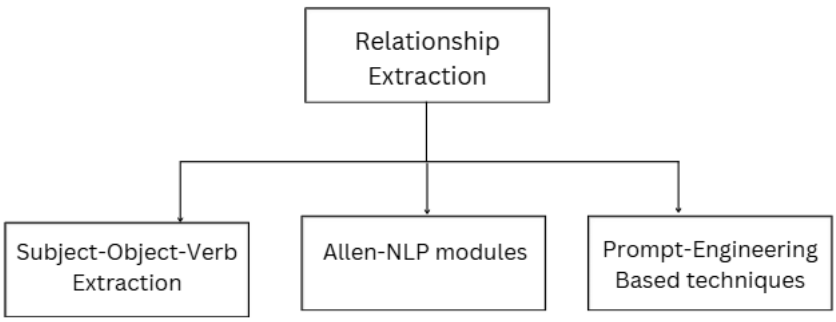

Figure 5.4: Comparison of Relationship Extraction Techniques

texts. Argument mining approaches aim to identify the structure of arguments in text, including relationships between claims and premises, which seems applicable to extracting the complex reasoning and causal relationships present in testimonies. This methodology promised to capture not just simple factual relationships but also the argumentative and explanatory structures through which survivors make sense of their experiences and convey understanding of historical events. However, according to our observation, the AllenNLP argument mining approach also proved inadequate for extracting relationships in Holocaust testimonies. The fundamental mismatch between argumentative discourse structures and narrative testimony structures meant that the models trained for argument mining could not effectively identify the types of relationships central to testimony content. Argument mining models are designed for texts where explicit reasoning patterns and claim-support relationships are linguistically marked, whereas testimonies express relationships through narrative progression, temporal sequencing, and implicit connections which are not according to the argumentative structures. Furthermore, the limited availability of annotated domain-specific data for training argument mining models, combined with the significant gap between existing argument mining corpora and testimony texts, prevented effective transfer learning of these methods. Therefore, the failure of both traditional rule-based and argument mining approaches demands further exploration of alternative methodologies that are more suited to the unique characteristics of Holocaust testimony texts.

As discussed in previous chapters, large language models (LLMs) have emerged as a promising solution for diverse NLP tasks. Their strength lies in an ability to understand context and handle linguistic variability without depending on extensive handcrafted rules or domain-specific training data. Trained on vast and diverse text corpora, contemporary LLMs possess broad linguistic knowledge and reasoning capabilities. This allows them to understand the semantic content of testimonies and identify entities and their relationships through contextual understanding, not just pattern matching. Through prompt engineering, they can also adapt to domain-specific terminology and relationship types, bypassing the need for large annotated datasets. Applying LLMs to relationship extraction in Holocaust testimonies reveals that deep semantic understanding is more critical than relying on syntactic patterns or argumentative structures. This is essential for interpreting meaning across varied linguistic forms, implicit references, and complex narrative structures. Furthermore, the process demonstrates that LLMs can be effectively guided by carefully designed prompts. These prompts specify the relationship types to extract, the output format for the results, and the contextual knowledge required to interpret the testimony language accurately.

The effectiveness of LLMs for extracting relationships from Holocaust testimonies derives from the limitations of previously tested traditional approaches. The model's ability to memorise long-range contexts enables them to resolve references and understand relationships that span multiple sentences, capturing the narrative coherence of the testimonies. Additionally, LLMs are flexible in handling varied linguistic and emotional features, accommodating the fragmented and unstructured sentences in oral testimony narratives. LLM's capacity for zero-shot and few-shot learning enables rapid adaptation to Holocaust-specific entities, events, and relationship types without requiring large annotated training datasets that would be highly costly and ethically complex to create. However, the length of Holocaust testimonies presents a significant practical challenge for LLM-based relationship extraction. Testimonies could range from several thousand to tens of thousands of

words, which raises the context window limitation of LLMs. Moreover, some studies have demonstrated that extended context capabilities encounter computational constraints and potential degradation in extraction quality when processing long documents (Xiaoxiong Wang and J. Hu, 2023). The model must maintain coherent knowledge throughout the entire testimony, which features a complex narrative structure where relationships and entities introduced early may be referenced later. Processing entire testimonies as single sentences becomes computationally costly while leading to missed relationships, inconsistent entity recognition across different sections, or degraded extraction quality in later portions of the testimony as the context becomes saturated with information. To address length constraints, it necessitates the development of segmentation strategies that divide testimonies into manageable chunks while preserving narrative coherence and contextual relationships.

The domain-specific nature of Holocaust testimonies requires that relationship extraction needed to be structured according to distinct thematic domains contained in survivor accounts. As illustrated in Figure 5.5, biographical relationships encompass personal and familial connections. Geographic relationships capture spatial information about locations where survivors lived, places they were deported to or from, camps and ghettos where they were imprisoned, hiding locations where they sought refuge, and routes of escape or forced marches they experienced. Historical and event-based relationships connect individuals to significant events, including experiences of persecution, acts of resistance or rescue, liberation circumstances, and post-war experiences. Each of these domains requires defined relationship types, and the definition must be comprehensive yet precise, accounting for the various ways that similar relationships might be expressed while maintaining consistency in the knowledge graph representation. For example, biographical relationship types might include family member relationships such as parent, child, sibling, spouse, and extended family, each with potential inverse relationships to ensure bidirectional connectivity in the graph. For geographic relationship types such as lived in, born in,

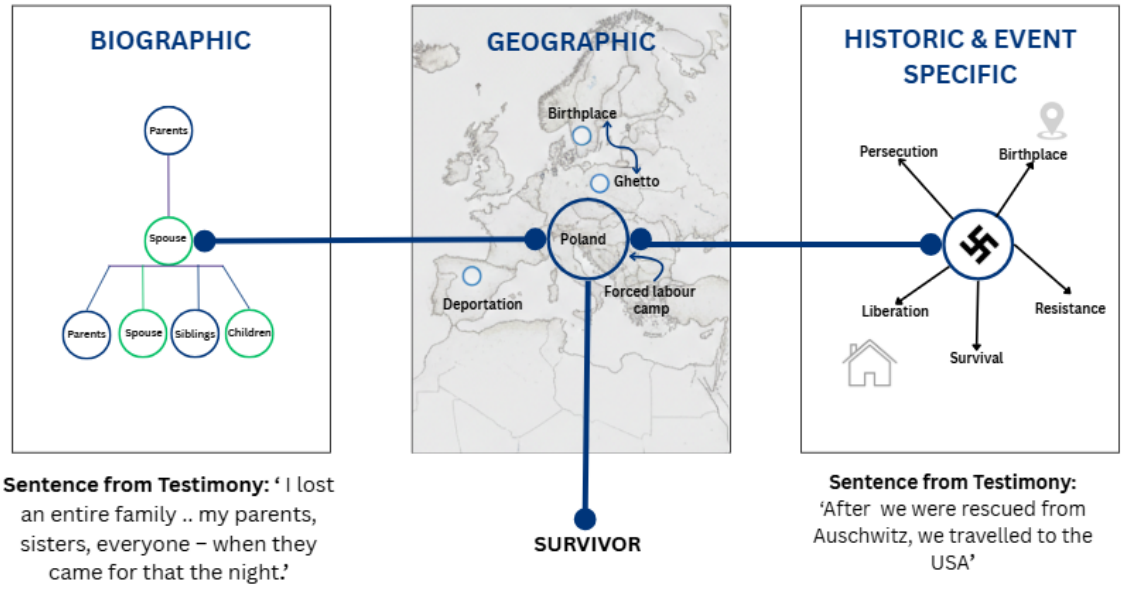

Figure 5.5: Mapping of Domain-Specific Relationship Types

deported to, imprisoned at, hidden in, escaped from, and liberated at, each carries specific semantic implications about the nature of the person-location connection. Historical relationships include what survivors have witnessed, participated in, survived, perished in, been rescued by, and been betrayed by, capturing the diverse ways individuals connected to events happened during the Holocaust period.

However, when recognising the above-discussed themes, the challenge of synonym recognition becomes critical when different testimonies or even different paragraphs of the same testimony use varied expressions to convey identical relationships. Survivors might describe the same type of relationship using different verbs/phrases, reflecting their own narrative styles, linguistic backgrounds, and the specific contexts based on their experiences. Domain adaptation for relationship extraction involves critical considerations specific to Holocaust testimonies. When employing prompt engineering techniques, the prompt might frame the extraction task to elicit accurate and appropriate responses, including clear instructions about what constitutes a relationship in the testimony context, examples of Holocaust-specific entities

and relationship types, guidance on handling temporal information and historical context, and instructions for maintaining sensitivity to the traumatic content while extracting factual information.

5.1.3 Result Evaluation

The model's relationship extraction performance was assessed against the manual annotation of 200 Holocaust testimonies, as described in Section 3, where human experts manually created ground-truth annotations. As given in Table 5.2, the LLM-based approach was applied to the same testimonies, and its outputs were compared against the manually annotated ground truth. Performance was measured using standard information extraction metrics: precision, recall, and F1-score. For biographical relationships, the proposed approach achieved a F1-score of 96.4%, indicating strong performance in identifying family connections and social relationships within survivor narratives. Career-based relationship extraction yielded a precision of F1-score 83.78%, and Holocaust event-based extraction obtained 85.4%. The results demonstrate that few-shot prompting can effectively extract structured relationship information from Holocaust testimonies, though challenges remain in handling ambiguous temporal references and complex familial structures characteristic of this historical context.

A qualitative error analysis was conducted on a random sample (50%) of false positives and false negatives to identify common failure patterns. According to our observations, the LLM-based approach was able to identify most hidden relationship patterns, even those that are ambiguous for humans to identify quickly. However, the following primary sources of error were observed, related to biographical relationships: (1) confusion between different types of family relationships when expressed ambiguously (e.g., interpreting `my aunt's daughter` as a direct relationship rather than recognising the cousin relationship), (2) difficulty with culturally specific relationship terms or Yiddish expressions that require domain knowledge, and (3) challenges in distinguishing between actual relationships and

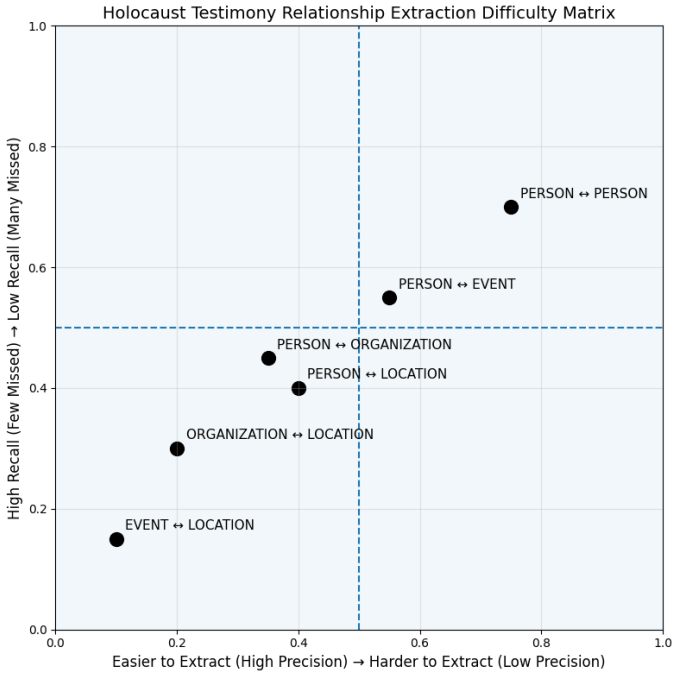

Figure 5.6: Relationship Extraction Difficulty Matrix by Entity Category

hypothetical or counterfactual statements (e.g., `my brother who died before I was born`). These error patterns highlight the need for enhanced contextual understanding and domain-specific knowledge in future iterations. According to our observations, Table 5.3 and Figure 5.6 analyse relationship extraction performance by most common entity pairs. Since no existing metric directly calculates the extraction difficulty, the difficulty matrix is calculated through the precision (ambiguities) and recall (missed detections) values, as illustrated in Figure 5.6. Extraction difficulty reflects the degree of ambiguity and detection failure. For most common relation types, position in a two-dimensional space is defined by precision-based difficulty $1 - P$ and recall-based difficulty $1 - R$, where difficulty is represented as the inverse of success rate.

5.2 Challenges of Relationship Extraction

Relationship extraction from Holocaust testimonies presents unique computational challenges that distinguish it from conventional information extraction tasks. In Holocaust testimonies, these challenges emerge from the intersection of traumatic narrative structure, historical complexity, linguistic variation, and ethical imperatives. Holocaust testimonies contain a non-chronological structure as survivors recall their memories without any order and scattered across multiple segments of testimony. This fragmentation complicates coreference resolution and requires contextual and discourse analysis to reconstruct relationship networks accurately. In testimonies, survivors use pronouns, nicknames, or contextual descriptions instead of formal names when referring to individuals. Phrases such as `my friend from the ghetto` or `the woman who helped us` require contextual inference to identify specific individuals, places and relationship types. The same person may be referred to differently across testimony segments, necessitating entity linking that accounts for these variations. Language switching is a common phenomenon in single testimonies, such as those in Yiddish, Hebrew, Polish, German, or other languages, alongside the primary language of the interview. In testimonies, relationship terms appear in any of these languages, and culturally specific kinship terms, which lack direct equivalents in the target language, require specialised lexicons and cross-lingual relationship ontologies.

Relationships in Holocaust testimonies are inherently temporal and personal when they are discussing the family members who were separated, killed, or disappeared. However, most of the existing relationship extraction methods could capture only the relationship types but not temporal validity, changes in status, and the occasions related to relationship dissolution. The challenge intensifies when testimonies provide uncertain or estimated timeframes. Moreover, for understanding relationships, additional historical knowledge of ghettos, camps, transport systems, and resistance networks is required. A `blockmate` relationship in Auschwitz carries different implications than in Theresienstadt. Survivors often possess incomplete

knowledge about the fate of family members and friends. Testimonies contain uncertainty markers, such as `I think`, `probably`, or `I never found out`, which highlight that relationship extraction must identify and preserve rather than treat all assertions as equally sure.

Extracting relationships from Holocaust testimonies requires interdisciplinary approaches that integrate NLP, cultural studies, and ethical reflection. The challenges outlined here underscore the need for careful, respectful, and historically informed computational approaches that serve the goals of Holocaust remembrance and education while honouring the complexity and humanity of survivor testimonies. Future work must balance technical innovation with deep engagement with the unique characteristics and ethical imperatives of this irreplaceable historical record.

5.3 Chapter Summary

This chapter discusses the methodology for modelling relationships within Holocaust survivor testimonies, highlighting that extracting meaningful connections from these narratives requires significant domain adaptation beyond standard NLP techniques. The proposed pipeline involves processing the testimonies, resolving complex coreferences to link pronouns through carefully engineered prompts employing LLMs, and extracting relationship triplets in different categories of relationship types. While evaluation shows strong performance, particularly for familial relationships, the process faces profound challenges inherent to the nature of the oral narratives. These include the non-chronological and fragmented narrative structure, linguistic variation including code-switching, the dynamic temporal nature of relationships, and the ethical imperative to handle traumatic content with sensitivity. Ultimately, the work demonstrates that effectively modelling these historical narratives requires an interdisciplinary approach, blending technical innovation with a deep contextual understanding and respect for the human experiences recorded.

Table 5.2: Performance Evaluation of LLM-Based Relationship Extraction against the human-annotated test set

Relationship Category	**Description / Focus**	**F1-Score (%)**
Biographical Relationships	Family connections and social relationships within survivor narratives.	96.40
Career-Based Relationships	Professional, occupational, or role-based connections mentioned in the testimonies.	83.78
Holocaust Event-Based Relationships	Relationships formed or defined in the context of specific Holocaust events (e.g., in camps, during hiding).	85.40

Table 5.3: Common Relationship Types Between Entities in Holocaust Testimonies

Relationship Between	**Frequent Relationship Types**
PERSON ↔ PERSON	child_of, parent_of, sibling_of, married_to, friend_of, hid_with
PERSON ↔ ORGANISATION	worked_for, was_prisoner_of, was_member_of, was_captured_by
PERSON ↔ LOCATION	was_born_in, lived_in, was_imprisoned_at, hid_in, fled_to
PERSON ↔ EVENT	witnessed, survived, was_victim_of, participated_in
ORGANISATION ↔ LOCATION	was_located_in, operated_at, had_subcamp_in
EVENT ↔ LOCATION	occurred_at, happened_in, was_centered_in

Chapter 6

Knowledge Extraction from Narrative Texts

For the survivor who chooses to testify, it is clear: his duty is to bear witness for the dead and for the living. He has no right to deprive future generations of a past that belongs to our collective memory.

Elie Wiesel (Survivor)

In the preceding chapters, we examined the information extraction techniques for extracting structured information from unstructured narrative texts. Chapter 4 introduced domain-specific NER, which enables the recognition of key named entities such as persons, locations, and domain-specific concepts. Subsequently, Chapter 5 explored relationship extraction methods, which identify the semantic connections between these entities, revealing the links within the narrative context. Together, these contributions established a framework for making Holocaust oral testimonies computationally processable and for extracting structured knowledge from them in a systematic way.

Yet extracting entities and relationships is not the same as understanding them. As defined in research question 02, this chapter addresses how those extracted fragments can be organised into a formal knowledge structure that supports reasoning, pattern recognition, and insight generation (basically the capabilities

that isolated extractions alone cannot provide). This question sits at the heart of building a coherent NLP framework for Holocaust narratives: one that moves beyond identifying *what* entities and relationships exist toward representing *how* they interconnect across testimonies, time periods, and geographies. This chapter therefore presents a comprehensive pipeline for constructing, organising, evaluating, and analysing a knowledge graph built from Holocaust testimony narratives, with the aim of transforming raw extraction outputs into a unified, queryable knowledge base that supports both historical research and computational analysis.

The gap between extracted information and actionable knowledge represents a fundamental challenge in natural language understanding. Raw outputs from information extraction pipelines would be formatted as entity mentions with type labels and relationship tuples, which suffer from several limitations:

- Fragmentation and Isolation: After individual extraction, it's challenging to trace connections across multiple testimonies or to understand the broader context in which entities and relationships operate. For example, a single concentration camp could be referenced by its official name, a euphemism, or its prisoners' slang (`Auschwitz III, Monowitz, or the Buna camp`), where these references remain disconnected.
- Lack of Semantic Structure: While relationship extraction identifies that two entities are connected, it does not provide an explanation of these connections within a richer semantic framework. Understanding that `the Nazis deported Jews from Warsaw to Treblinka` is informative, but recognising it as part of a broader systematic extermination pattern which was linked to organisational, ideological, and logistical structures offers far deeper analytical insight.
- Limited Reasoning Capability: Isolated extractions do not support inferential reasoning. If we know `The Gestapo operated in France` and `The Gestapo reported to the RSHA in Berlin`, we cannot directly infer without a structured representation that `The RSHA directly controlled all operations`

`in France`.

- Scalability and Reusability Concerns: As the volume of data grows, managing isolated extractions becomes impractical, and structured knowledge representation enables efficient storage, retrieval, and updating of information, while also facilitating knowledge transfer across different applications and domains.

6.1 Information Extraction to Structured Knowledge

This section discusses the systematic approach of transforming extracted entities and relationships into structured, queryable, and analysable knowledge representations: knowledge graphs and ontologies. Knowledge graphs provide a flexible structure for representing extracted knowledge as interconnected networks of entities and relationships. In a knowledge graph, entities become nodes and relationships become edges. This graph representation allows capturing the network structure of the content by connecting events, actors, and concepts. Moreover, knowledge graphs help integrate information from multiple sources, cross-reference entities, and support complex queries that traverse multiple relationship paths.

Alternative representation approaches were considered but were not fully suitable for the analytical and interpretative goals of this study. Relational databases provide clear schemas, strong data integrity, and efficient storage. However, they depend on fixed table structures that are difficult to adapt to evolving historical categories and complex many-to-many relationships. Representing detailed event structures, overlapping time periods, or uncertain classifications often requires additional linking tables and schema modifications, which reduces flexibility and increases structural complexity. Document-based or vector-based representations (such as embedding spaces) are useful for large-scale retrieval and semantic similarity search. However, they store knowledge as numerical representations rather than as clearly defined relationships between entities. This makes the structure less transparent and limits

explainable querying and logical reasoning, which are particularly important in historically sensitive contexts. Property graph models provide flexible connections between entities, but, without a formal semantic layer, they lack clearly defined class hierarchies, property constraints, and shared vocabularies. As a result, they offer limited support for consistency checking, automated reasoning, and interoperability with external systems. In a domain that requires conceptual clarity, traceability, and formal validation, the absence of an explicit ontological framework is a significant limitation.

Ontologies provide the formal semantic structure that provides a meaning to knowledge graphs. While a knowledge graph represents specific instances, such as `Auschwitz was a camp`, an ontology defines the formal structure of classes, properties, and constraints that govern these instances (e.g., Auschwitz was an extermination camp, which is a subclass of camp, which is located in a place, and properties with operational periods and liberated by an army). Ontologies facilitate consistency checking, support automated reasoning, and ensure that extracted knowledge attaches to the domain-specific information.

Together, knowledge graphs and ontologies transform extracted information into a rich, formal knowledge base that supports advanced analytical capabilities. This structured format allows researchers to further explore simple fact retrieval toward refined knowledge discovery, including pattern identification, anomaly detection, and predictive reasoning. The approach presented in this chapter follows a systematic pipeline that transforms narrative texts into analysed knowledge structures which are visualised as given in Figure 6.1.

6.1.1 Fundamentals of Knowledge Graph

Knowledge graphs have emerged as a method of representing structured knowledge extracted from unstructured data sources, providing a formal, direct and labelled graph representation of entities and their connections.

A graph G can be defined as a tuple :

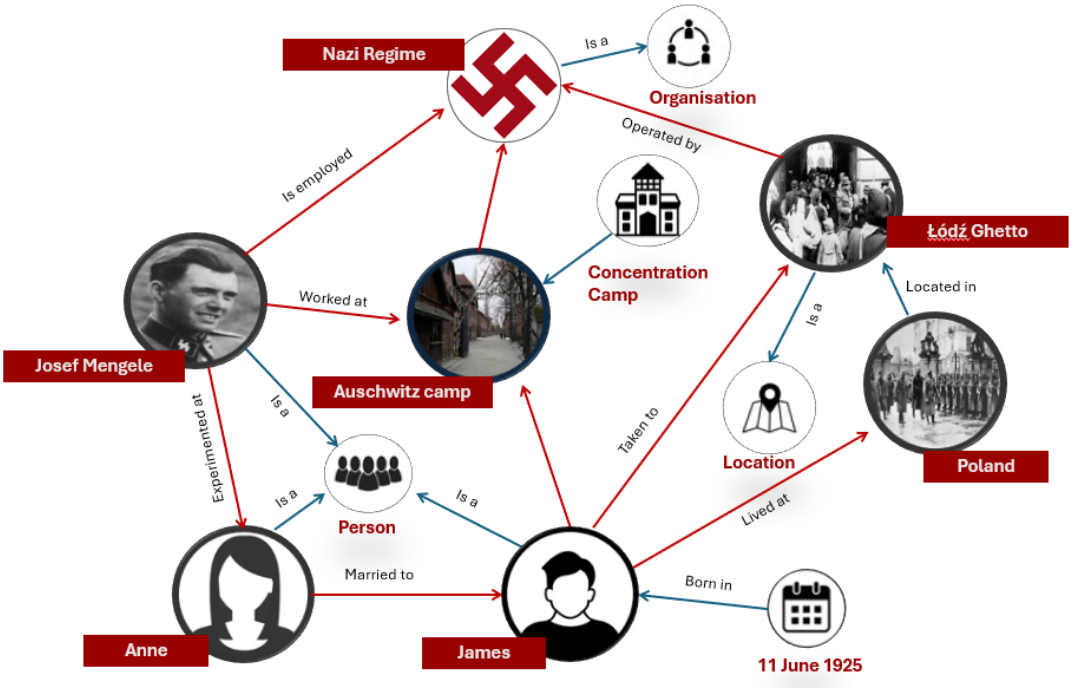

Figure 6.1: Pictorial visualisation of proposed knowledge graph structure

$$G = (V, E, L_V, L_E, \phi)$$

where:

- V is a set of vertices (nodes) representing entities
- $E \subseteq V \times V$ is a set of directed edges representing relationships between entities
- L_V is a set of labels for vertices (entity types)
- L_E is a set of labels for edges (relationship types)
- $\phi : V \cup E \to P$ is a function mapping vertices and edges to sets of properties

The fundamental components of a knowledge graph include:

- Nodes (Entities): Nodes represent entities extracted from text, including concrete objects, abstract concepts, events, or any identifiable item of interest. Each node consists of a unique identifier, a type or class label and a set of properties.

- Edges (Relationships): Edges represent directed relationships between entities. Each edge connects a source node (subject) to a target node (object) and carries a relationship type that characterises the nature of the connection. Edges may also have properties that provide additional context about the relationship, such as temporal information, certainty scores, etc.

- Properties (Attributes): Properties are key-value pairs attached with nodes and edges to capture descriptive information beyond simple categorical labels. Common attribute types are references to external resources (URLs, database IDs), provenance information (source documents, extraction confidence), and literal/temporal markers (strings, numbers, dates). Moreover, properties allow a model to represent both categorical information (types and relationship labels) and continuous or complex attributes in parallel.

Beyond simply consolidating information, the graph structure itself carries semantic meaning that enables advanced analytical capabilities. Unlike traditional relational databases, where relationships are implicit in table joins, knowledge graphs make relationships explicit. This structure allows for path-based reasoning (traverse between multiple edges to discover indirect connections), network analysis (apply graph algorithms to identify central entities, communities, or structural patterns), flexible schema (add new entity types or relationship types without restructuring existing data), and integration (merge knowledge from multiple sources by connecting nodes that represent the same entities).

Understanding distinct and domain-specific information is crucial for the accuracy and reliability of the knowledge graphs. A properly formatted knowledge graph consists of unified entity representation, which resolves all mentions of the same entity to a single node in the graph. This node accumulates all information about the entity from any source document. Relationships become edges of the knowledge graph by connecting entity nodes and creating a network that persists beyond individual documents. Additionally, detailed information from multiple sources is consolidated into a single, coherent structure, answering complex questions

by traversing graph paths with reasoning capability. This structural formulation enables the performance of graph-based querying, multi-hop reasoning, centrality analysis, community detection, and complex pattern mining, all of which are impractical with raw, isolated extractions.

6.1.2 Graph Construction Process

Knowledge graph construction from Holocaust testimonies requires specialised approaches that account for the unique characteristics of testimonial discourse, historical trauma narratives, and the critical importance of accuracy and respect in representing survivor accounts. This section explains the procedure that was adapted for constructing knowledge graphs from Holocaust testimonies by addressing domain-specific challenges and the ethical obligations inherent in representing survivor experiences.

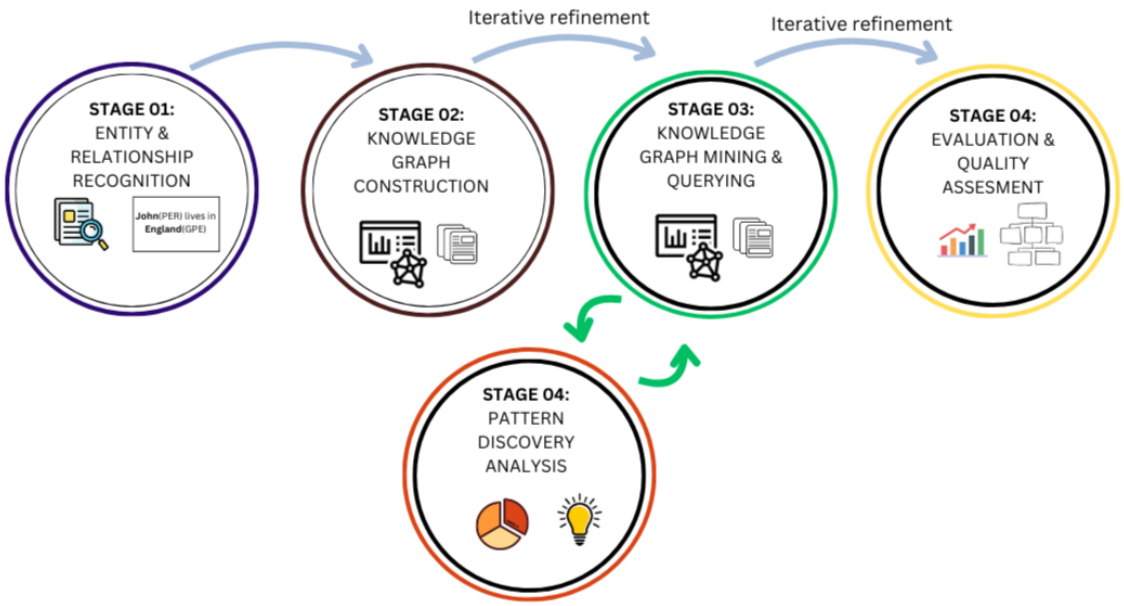

Figure 6.2: Knowledge Extraction Pipeline

As illustrated in Figure 6.2, the approach presented in this section follows a systematic pipeline that transforms narrative texts into analysed knowledge

structures through five interconnected stages:

1. Stage 1: Entity and Relationship Extraction builds upon the techniques established in previous chapters. The NER methods were used to identify and classify the domain-specific entities, while relationship extraction detects connections between pairs of entities, which will be the raw material for knowledge graph construction: a collection of entity mentions with types and a set of relationship instances.
2. Stage 2: Knowledge Graph Construction transforms extracted entities and relationships into a unified graph structure (RDF format). Entity resolution was handled to consolidate different mentions of the same entity (pronoun resolution). Followed up by conducting relationship normalisation to map diverse expressions to a standardised format. And finally, a graph was constructed to create a coherent network representation. The output is a knowledge graph where nodes represent entities, edges represent relationships, and properties capture attributes and metadata.
3. Stage 3: Knowledge Graph Mining and Querying provides systematic interaction of structured knowledge. This stage involves the development of a query interface of SPARQL for structured retrieval of facts, entity-centric information, and complex subgraph patterns. Graph mining techniques are utilised to calculate centrality metrics that identify influential entities, uncover community structures, and conduct path analysis to trace interconnections between entities.
4. Stage 4: Pattern discovery and similarity analysis examine the structure and content of the knowledge graph to identify recurring patterns, behaviours, and regularities. Similarity metrics were employed to identify patterns consistent across various sections of the graph. The output includes a comprehensive list of identified patterns, detailing their frequency and distribution, similarity clusters that categorise related entities or subgraphs, and domain-specific insights that transform structural findings into meaningful interpretations.

5. Stage 05: Evaluation and Quality Assessment measures the quality of the knowledge graph. The evaluation stage proceeds with two dimensions: intrinsic quality metrics were employed to assess the internal consistency and completeness of the graph, while extrinsic evaluation analyses the performance on downstream tasks, and comparative analysis benchmarks the knowledge representation against gold standards.

6.1.2.1 Knowledge Graph Construction

The construction of a knowledge graph from Holocaust testimonies involves several critical stages, each addressing specific challenges in transforming unstructured narrative text into structured, queryable knowledge.

Entity and Relationship Extraction

As discussed in previous sections, entities and relationships were extracted using a combination of automatic methodologies. The extraction pipeline integrates multiple NLP techniques, such as NER, to identify and link entities across testimony narratives.

Coreference and Pronoun Resolution

The impact of unresolved pronouns on the knowledge base structure is significant and measurable. Without pronoun resolution, the RDF knowledge base would contain a triple stating that an entity called `He` worked as a doctor, and the knowledge base would then create a separate person entry for `He` with its own set of attributes. Meanwhile, the knowledge base would also have `Perets Barskiy` (who is related to him) as a separate person with entirely different attributes. The result is having two duplicate entities in the knowledge base representing the same historical person. This duplication creates multiple problems: first, queries become unreliable because searching for information about `Perets Barskiy` will miss facts stored under `He`; second, statistical analyses of the knowledge base become inaccurate because it appears to have more entities than the actual number of entities; third, it is hard to create family trees and reconstruct relationships with incomplete and

fragmented results. Therefore, pronoun resolution is a critical component in the process of building a knowledge graph. As the section 5.1.1.1 describes, LLM-based approaches were utilised for coreference resolution in the Holocaust testimonies. According to our observation, the majority of narratives were conveyed from a third-person viewpoint with an unstructured format, complicating pronoun and coreference resolution. Although various rule-based methodologies such as part-of-speech tagging are being employed for small sets of samples and tested, none effectively address resolving the complex familial relationships inherent in the data.

Duplicate detection and Data validation

At the beginning of the stage, all the unrefined textual triples required for the structured knowledge graph were being preprocessed and cleaned. This step ensures the reliability, consistency, and integrity of the final graph by systematically mitigating the noise inherent to narrative testimonies. The primary issue is the identification of duplicate entities and resolution of them. Holocaust testimonies refer to the same individuals, locations, and events using different phrasings, resulting in multiple, inconsistent representations within the extracted triples. Therefore, this stage was addressed at multiple levels:

- Elimination of Duplicate Entities: Identical triples, resulting from repeated references in a testimony, are easily identifiable and merged into a singular, unique declaration.

- Near-Duplicate Resolution: More complicated to identify are near-duplicates, where the same real-world entity or relationship is expressed with linguistic variation. For example, `Warsaw Ghetto`, `the ghetto in Warsaw`, and `Ghetto Warszawa` all reference the same landmark.

To address the above-mentioned steps, a fuzzy matching approach was adapted, utilising string similarity measures such as Levenshtein distance (for character-level modifications) and Jaccard similarity (for token-based overlap), followed by clustering the ones with the same meaning. This cleaning phase is essential for

creating an organised and de-noised set of triples required for the relationship normalisation and knowledge graph structuring.

Data validation and quality checks

Data validation and quality checks ensure the knowledge graph is structurally sound and historically accurate. Structural validation was conducted to verify that all triples are in the same format, with each triple containing three well-formed components (subject, predicate, object), predicates were related to the context; and entities were properly typed according to the classification schema. Moreover, manual assessment was conducted to examine whether relationships make logical sense given the entity types, verifying that each relationship connects the Holocaust survivor to the camp or ghetto rather than to the Nazi officer or guard. As a part of this validation process, historical validation checks were conducted to verify extracted information such as concentration camps, ghettos, and other locations existed during the periods mentioned in the established Holocaust terminologies (EHRI and USHMM vocabulary lists).

Entity Linking

Entity linking addresses the challenge of connecting entities extracted from Holocaust testimonies to authoritative external knowledge bases. In this study, entity-linking process maps identified entity types, such as persons, locations, and organisations, to the corresponding entries in external resources, such as Holocaust databases and Wikipedia. After selecting possible external Holocaust databases (USHMM, EHRI, Centropa and Wikipedia), meta information (URLs related to the testimony) has been combined for each triple after the validation of that external knowledge base containing the information. By incorporating this information as meta-information, the process enhances the interpretability, discoverability, and preservation of Holocaust testimonies. The primary goal of entity linking is to include meta information in the URI format related to the triples. It allows researchers to cross-reference information, validate facts, and uncover connections between people, places, and events across different survivor accounts. Furthermore,

this approach could help to combat misinformation by anchoring personal memories to verified historical data.

Relationship Normalisation

Relationship normalisation standardises the way relationships between entities appear the same on the knowledge graph. For example, `born_in` and `born_on` are two similar associations that could represent distinct but related information. On the surface, the issue might seem like a minor inconsistency, but the failure to normalise relationships creates serious challenges for querying and reasoning. Normalisation issues are common because natural languages contain the same semantic relationships in many different ways. The statements `Freida Borschevskaya` was born in `Romny` and `Freida Borschevskaya was born in 1888` both clarify when and where `Freida` was born, but different predicates were used to connect different object types. Without proper normalisation, the knowledge base would have multiple representations of the same relationship, or relationships that combine might be kept separate, leading to incomplete and inconsistent knowledge.

Therefore, relationship normalisation is critical for several practical reasons. In some cases, unnormalised relationships cause query failures; for example, if a SPARQL query tries to "find all people born in a certain location," it searches for relationships with the predicate `born_in`. If some birth location information is stored with the predicate `born_in` and other location information is stored with different predicates, such as `was_born_at` or `birthplace`, the query misses the relevant results and returns incomplete answers because it only searches for one predicate form. Furthermore, knowledge bases with unnormalised relationships are more complicated to understand and maintain. If multiple predicates contain the same meaning, future users of the knowledge base will be unaware of which predicate to use when querying. This confusion leads to errors and misuse of the knowledge base. Because of that, relationship normalisation is necessary for data integration in scenarios such as merging the testimony knowledge base with other historical databases, where the relationships need to be standardised so that data

from different sources can be properly integrated.

RDF graph construction

Following relationship normalisation, the cleaned and standardised triples are transformed into an RDF knowledge graph structure. Each entity identified was assigned a unique URI that serves as a persistent identifier within the graph. These URIs contain a consistent naming convention that combines the entity type and a unique identifier, ensuring that each entity can be unambiguously referenced and linked to the knowledge base. The normalised predicates are mapped to defined relationship types within the graph schema, implementing a controlled vocabulary of relationship types that maintain semantic consistency throughout the knowledge graph.

RDF triples are produced in the Turtle (.ttl) serialisation format during the graph construction process. The GraphDB database is then used to load the .ttl file resulting in an RDF graph, which offers efficient storage, indexing, and SPARQL query capabilities. This structured representation enables systematic querying of Holocaust testimonies, allowing one to trace relationships across multiple testimonies. The RDF format facilitates integration with external knowledge bases through the entity linking URIs established in the previous stage, creating connections between the testimony knowledge graph and Holocaust databases.

Graph Storage and Indexing

The proposed knowledge graph requires proper storage and indexing mechanisms to support complex queries and analytical operations. The graph is stored in a dedicated graph database (GraphDB) that provides native support for traversing relationships and executing pattern-matching queries. Indexing strategies are able to optimise query performance, including entity-based indices that enable rapid retrieval of specific persons, locations, or events, and relationship-based indices that facilitate efficient traversal of connections between entities. The graphs' storage maintains both the graph structure (RDF format) and associated metadata, including additional information that traces each triple back to its source testimony.

Moreover, to handle the scale of Holocaust testimony collections, GraphDB's storage supports updates, allowing new testimonies to be processed and integrated into the existing knowledge graph without requiring complete reconstruction. Additionally, GraphDB's robust SPARQL query support, built-in visual exploration tools, scalability for large knowledge graphs, and comprehensive reasoning features were the primary reasons for its selection as the triplestore. The GraphDB has multiple indexing strategies to optimise query performance: subject-predicate-object (SPO) indices enable rapid retrieval of all information about a specific entity, predicate-object-subject (POS) indices facilitate efficient searches for entities sharing particular relationships, and object-subject-predicate (OSP) indices support reverse relationship traversal. These indices are maintained by the triple store as new triples are added or modified, providing consistent query performance when expanding the knowledge graph with new testimonies. Furthermore, GraphDB's visual graph interface provides interactive illustrations of the knowledge graph to visually navigate entity relationships, identify connection patterns, and discover implicit links between survivors. The indexed triple store provides the foundation for the knowledge graph mining and querying stage, enabling complex SPARQL queries that would be computationally infeasible without proper indexing structures. The next section would discuss the graph mining and querying approaches conducted by using SPARQL.

6.1.2.2 Knowledge Graph Mining and Querying

The knowledge graph created from Holocaust testimonies enables organised searching and analysis of information which was difficult to find using regular keyword searches. This stage involves two complementary approaches: 1) structured querying through SPARQL to retrieve specific facts and entity-centric information, and 2) graph analytics to uncover implicit connections, measure entity importance, and identify thematic groupings within the Holocaust testimonies. These features transform the knowledge graph into a powerful research tool that supports targeted

domain-specific information retrieval and exploratory discovery.

SPARQL Query Interface

The knowledge graph is queried using SPARQL (SPARQL Protocol and RDF Query Language), a W3C standard query language designed specifically for RDF data. SPARQL provides researchers with the ability to create precise queries that traverse the graph structure, retrieve entity attributes, and discover relationships across multiple testimonies. The proposed query interface supports several categories of information retrieval relevant to this study:

- **Entity-centric queries**: Retrieve all information known about a specific person, location, or event. For example, finding all testimonies that mention a concentration camp.

```
PREFIX ex: <http://natrix.org/holocaust/>
SELECT ?predicate ?object
WHERE {
ex:Hanna_Ferber ?predicate ?object
}
```

- **Relationship queries**: Identify entities connected by specific relationship types. Examples include finding all PERSON-CAMP relationships in the corpus, identifying survivors who were imprisoned in the same camp, or tracing deportation routes to concentration camps.

```
PREFIX ex: <http://natrix.org/holocaust/>
SELECT ?person ?camp
WHERE {
?person ex:imprisoned_in ?camp.
}
```

- **Temporal queries**: Filter and retrieve information based on temporal constraints, such as identifying all events that occurred during a specific time period or finding survivors born in a particular decade.

```
PREFIX ex: <http://natrix.org/holocaust/>
SELECT ?person
WHERE {
?person ex:born_in ?date .
FILTER(CONTAINS(STR(?date), "1919"))
}
```

- **Multi-hop path queries**: Discover indirect connections between entities by traversing multiple relationship edges. This enables researchers to trace family connections across generations, identify shared experiences between survivors who never met, or map migration patterns through multiple locations.

```
PREFIX ex: <http://natrix.org/holocaust/>
SELECT ?other
WHERE {
ex:Hanna_Ferber (ex:born_in | ex:imprisoned_in) ?other .
}
```

Path Analysis and Connection Discovery

Path analysis explores the indirect connections between entities by traversing multiple relationship edges in the knowledge graph. Its application to the Holocaust domain enables researchers to formally explore and reveal implicit networks that exist connecting individuals through intermediary experiences, witnesses, or locations. Technically, this was achieved through SPARQL property path queries, where the path length is parameterised to manage the trade-off between the scope of discovery and computational complexity.

- **Shared location sequences**: Finding survivors who followed similar deportation or escape routes, even if they were there at different times. For example, identifying all paths connecting survivors from a particular ghetto to a specific concentration camp reveals the common trajectories of persecution.

```
PREFIX ex: <http://natrix.org/holocaust/>
SELECT ?person ?birthPlace ?livedPlace ?camp
WHERE {
?person a ex:PERSON .

?person ex:born_in ?birthPlace .
?person ex:lived_in ?livedPlace .
?person ex:imprisoned_in ?camp .
}
ORDER BY ?birthPlace ?livedPlace ?camp ?person
```

- **Family networks across testimonies**: Connecting mentions of the same family members across different testimonies allows reconstruction of family histories that no single testimony fully documents. Cross-testimony family networks were reconstructed by identifying individuals connected via familial relationships who were referenced across independent survivor narratives. For example, if one survivor mentions their uncle and another testimony from a different witness mentions the same person (uncle), path analysis can link these references.

```
PREFIX ex: <http://natrix.org/holocaust/>
SELECT ?person1 ?relation ?person2 ?witness
WHERE {
?witness ex:mentions ?person1 .
?person1 ?relation ?person2 .
```

```
FILTER(?relation IN (
    ex:has_parent,
    ex:has_child,
    ex:has_sibling,
    ex:married_to
))
}
```

- **Witness co-occurrence**: Identifying which survivors mention the same people, places, or events, even when they don't directly mention each other. This creates a network of shared witnessing that reveals which experiences were collectively documented.

```
PREFIX ex: <http://natrix.org/holocaust/>
SELECT ?w1 ?w2 ?camp
WHERE {
    ?w1 ex:imprisoned_in ?camp .
    ?w2 ex:imprisoned_in ?camp .

    FILTER(?w1 != ?w2)
}
ORDER BY ?camp
```

Graph Analytics and Metrics

Beyond direct querying, graph analytics techniques compute structural properties that reveal the relative importance and connectivity of entities within the knowledge graph. After extracting all related information from GraphDB, NetworkX was used to compute these centrality metrics, which provide quantitative rankings of

entity importance based on different connectivity patterns. NetworkX was selected for its superior interactive visualisation capabilities compared to GraphDB's native options.

Degree Centrality: measures the number of direct connections an entity has within the graph. In the context of Holocaust testimonies, a high degree of centrality indicates entities (persons, locations, and organisations) are frequently mentioned across multiple testimonies. This metric was used to identify the most documented and interconnected elements of the Holocaust experience.

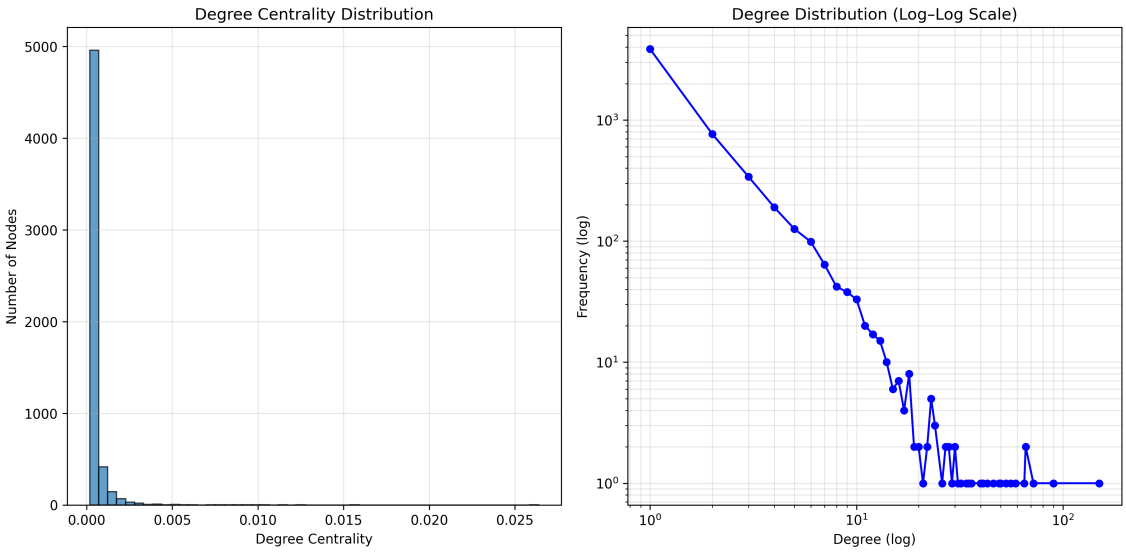

Figure 6.3: Degree of Centrality.[1]

Figure 6.3 presents the degree centrality distribution of the constructed knowledge graph. The distribution presents a heavily right-skewed pattern, with approximately 95% of nodes having degree centrality below 0.001. This indicates a highly centralised network structure where the majority of entities have few connections, while a small subset of hub nodes possess significantly higher connectivity. The log-log plot illustrates a power-law distribution, evidenced by the approximately linear relationship between degree and frequency on logarithmic scales. The power-

[1]This visualisation is valid only for the dataset extracted from the Centropa archive.

law behaviour indicates that the knowledge graph follows typical patterns observed in semantic networks, where certain entities (frequently referenced entities) have more connections than others. For example, major concentration camps such as `Auschwitz` have a high degree of centrality because multiple survivors mention imprisonment there, and the camp is associated with various relationship types such as `imprisoned_in, deported_to, liberated_from`.

Betweenness Centrality identifies entities that serve as bridges between different parts of the graph connecting other entities. In a knowledge graph, high betweenness centrality indicates transit camps or ghettos through which many survivors passed or individuals who connect different survivor communities. This metric is particularly valuable for understanding deportation routes and identifying locations that served as critical centres in the persecution process.

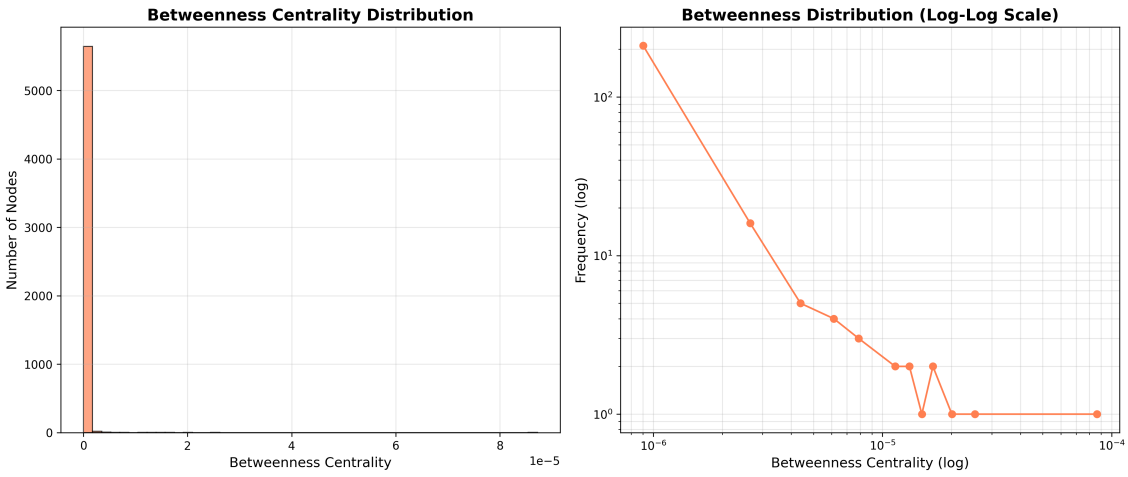

Figure 6.4: Betweenness Centrality.[2]

Figure 6.4 presents the betweenness centrality distribution of the knowledge graph. Similar to degree centrality, the distribution exhibits extreme right skewness, with the majority of nodes (>95%) having near-zero betweenness centrality. The log-log plot confirms power-law characteristics, consistent with scale-free network

[2]This visualisation is valid only for the dataset extracted from the Centropa archive.

properties. The correlation between degree and betweenness centrality is moderate (r = 0.440, Figure 6.4), indicating highly connected nodes tend to have higher betweenness. This suggests a community structure where hub nodes exist within dense clusters while bridge nodes with relatively low degrees connect disparate parts of the graph. According to the experiments, entities such as `Bitola`, `Gestapo`, `Russians`, and `Bulgarians` act as the bridge nodes with high betweenness-to-degree ratios, highlighting their role as critical connectors between different semantic communities in the knowledge graph. These nodes represent the narrative structure, linking geographic, political, and social aspects in Holocaust testimonies.

Closeness Centrality measures how quickly an entity can reach all other entities in the graph through relationship paths. Entities with high closeness centrality are well-connected to the broader network and can be thought of as central to the whole testimony. This metric helps identify geographically central locations and persons whose experiences intersect with many others' stories.

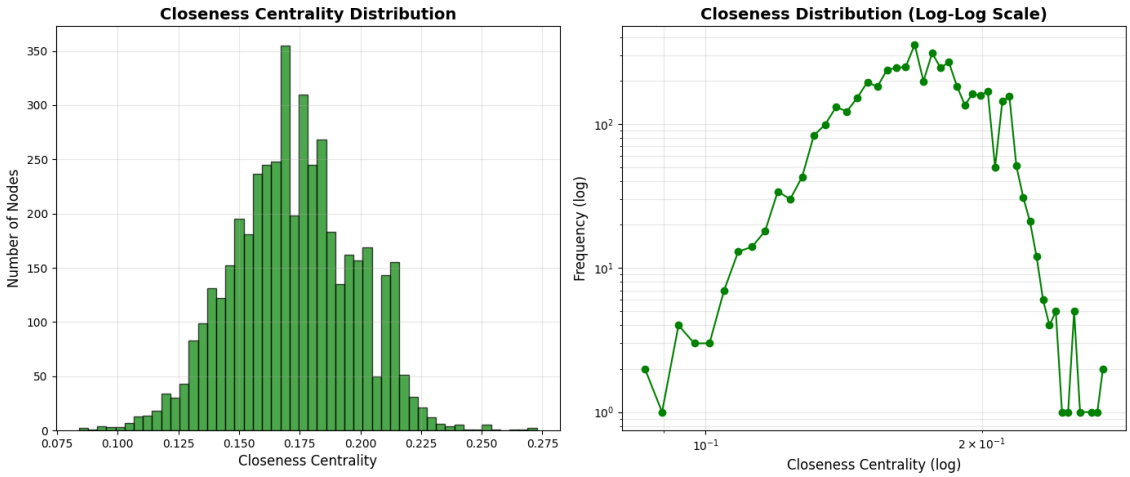

Figure 6.5: Closeness Centrality.[3]

Figure 6.5 presents the closeness centrality distribution, which illustrates a different pattern from degree and betweenness centrality. Unlike the extreme right-

[3]This visualisation is valid only for the dataset extracted from the Centropa archive.

skewed distributions observed in degree and betweenness, closeness centrality follows an approximately normal distribution centred around 0.17, with values ranging from 0.08 to 0.27. This indicates that most entities in the knowledge graph maintain relatively uniform shortest path distances to other entities, suggesting good structural cohesion within the main connected component. The correlation analysis illustrated in Figure 6.6 shows the distinct relationships between centrality measures. While degree and betweenness present a very strong correlation ($r = 0.921$), indicating that hubs also serve as primary bridges in the network, both show only a weak-to-moderate correlation with closeness ($r = 0.383$ and $r = 0.309$, respectively). This pattern reveals major conceptual themes (`Israel`, `family`, `Kiev`, `war`), as highly connected hubs. Meanwhile, closeness centrality identifies additional, centrally positioned nodes that efficiently bridge disparate parts of the narrative space.

The most central nodes by closeness centrality include both major geographic locations (`Israel, Kiev, Odessa, Auschwitz`) and core relational concepts (`family, parents`), representing semantic features that efficiently bridge different testimony narratives. In contrast, peripheral nodes with low closeness (< 0.10) represent highly specific details (`Jewish ghetto in Shanghai (a visa-free location in China where Jews could flee between 1933-1941)`, specific individuals with limited mentions) that appear in isolated testimonies, reflecting unique experiences not widely shared across the corpus.

Community Detection

Community detection algorithms identify clusters which were closely linked entities within the knowledge graph, revealing thematic groupings. Community detection was performed using the Louvain modularity optimisation method, which has identified 47 distinct communities within the main connected component (4,537 nodes), achieving a high modularity score of 0.7923. This indicates a strong community structure where entities cluster into coherent thematic groups with dense

[4]This visualisation is valid only for the dataset extracted from the Centropa archive.

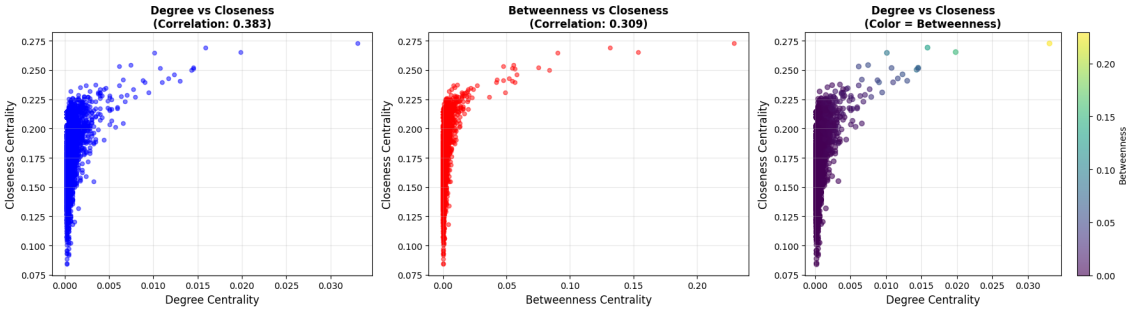

Figure 6.6: Comparing centrality measures.[4]

internal connections and sparse inter-community links.

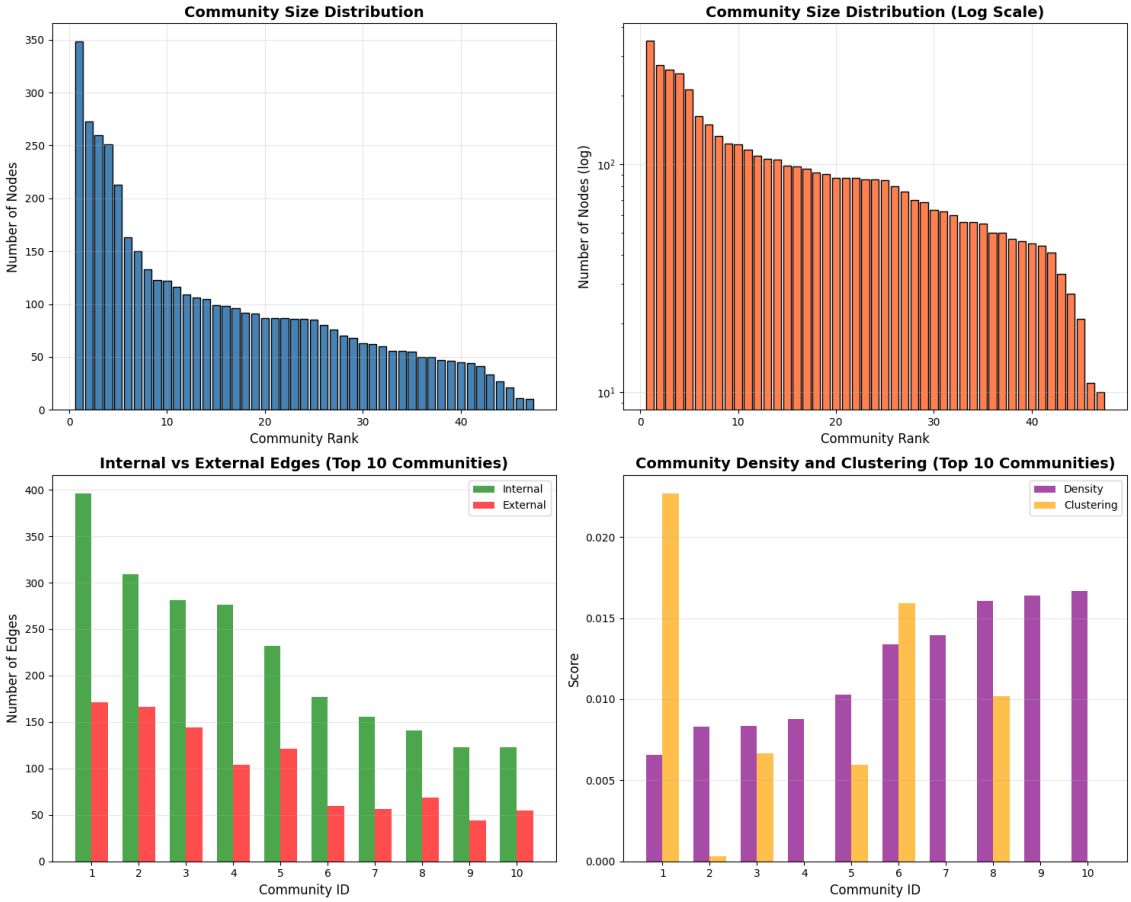

Figure 6.7: Overall community distribution

According to our observation, community size distribution presents substantial heterogeneity: the largest community contains 348 nodes (7.67% of the main component), while the smallest contains 10 nodes, with an average community size of 96.5 nodes. The top 10 communities collectively represent 43.9% of Figure 6.7, indicating that testimony entities concentrate into several major thematic clusters rather than dividing uniformly across many small groups. The log-scale distribution reveals a long tail of smaller communities, reflecting the diverse range of locations, individuals, and events documented across testimonies. Density and clustering metrics show that individual communities maintain low internal density (0.0066-0.0167) but exhibit clustering coefficients ranging from near-zero to 0.0227, higher than the graph-level average (0.0044). This indicates that communities are not fully connected internally, but they contain localised substructures that increase local clustering relative to the global network structure.

Temporal patterns in the graph show that communities span multiple historical periods rather than clustering by era. As shown in Figure 6.8 overall temporal distribution shows concentration around 1939-1945 (27.3% of year mentions), with substantial pre-war (1900-1938: 33.3%) and post-war (1946-2000: 28.5%) coverage, indicating that testimonies situate wartime persecution within comprehensive life narratives rather than focusing exclusively on the Holocaust period.

Figure 6.9 illustrates the inter-community connectivity that forms a dense network where the top 10 communities maintain 110-342 external connections each. The strongest inter-community links include Community 1 & Community 3 (21 connections) via `Israel, Jews, Germans` and Community 2 & Community 5 (20 connections) via `Kiev, war, family`, demonstrating that communities bridge through shared geographic and collective entity references. Bridge node analysis identifies 842 entities connecting multiple communities, with `Israel` (33 communities), `Kiev` (25 communities), and `Odessa` (22 communities) serving as primary inter-community connectors. These geographic hubs integrate diverse local narratives into a coherent transnational network.

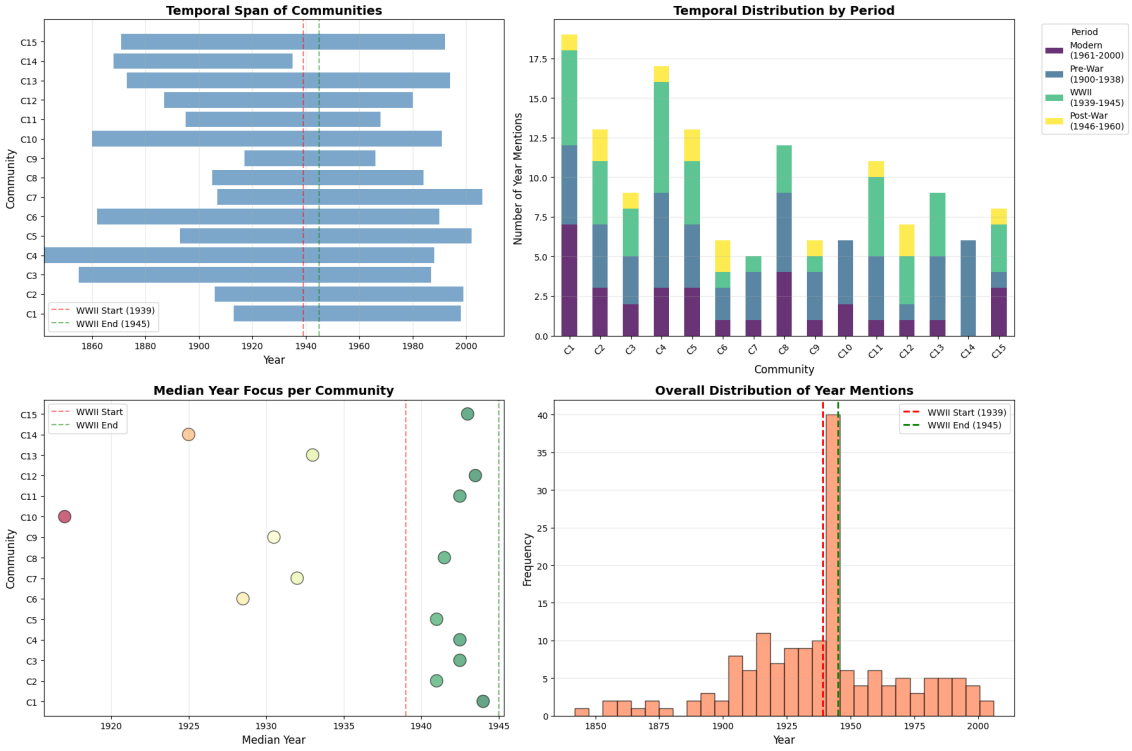

Figure 6.8: Temporal patterns in the community distribution

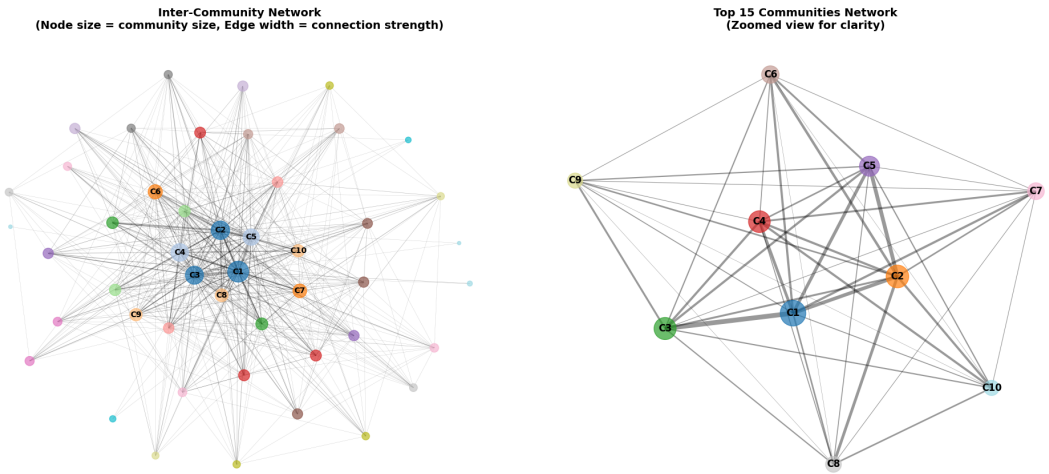

Figure 6.9: Intercommunity network

Moreover, Figure 6.10 illustrates thematic and distinct semantic pattern analysis of the top communities:

- Community 1 (348 nodes): The largest cluster centres on collective perpetrator-victim dynamics, with `Germans` and `Jews` as top hubs, alongside individual witnesses such as `Alexander Gajdos` and the military entity `army`. Internal density is 0.0066 with 396 internal edges versus 171 external connections.

- Community 2 (273 nodes): A geographic-familial cluster dominated by `Kiev` (the second-highest degree hub in the entire graph) and `family`, with strong connections to wartime evacuation themes. This community exhibits higher connectivity to other clusters (332 inter-community edges), serving as a geographic bridge.

- Community 3 (260 nodes): Post-war emigration and survival cluster centred on `Israel` (the graph's highest-degree hub), `ghetto`, and `parents`. Its 288 inter-community connections reflect Israel's role as a common destination linking diverse pre-war origins.

- Community 4 (251 nodes): The `Theresienstadt/Czech` persecution cluster focused on `Terezin, Holocaust`, and `Prague` representing the Central European deportation system.

- Communities 7-8: Auschwitz-centred persecution (Community 7, 150 nodes) and Moscow-centred Soviet experience (Community 8, 133 nodes), representing geographically and thematically distinct Holocaust narratives.

According to the Table 6.1 having a high modularity score (0.79) confirms that the graph exhibits a strong natural community structure that aligns with historical and thematic patterns rather than being an artefact of arbitrary clustering. Communities correspond to recognisable Holocaust narrative types: perpetrator-victim dynamics (C1), geographic centres (C2, C8), emigration destinations (C3), and specific persecution infrastructures (C4, C7). The moderate community sizes

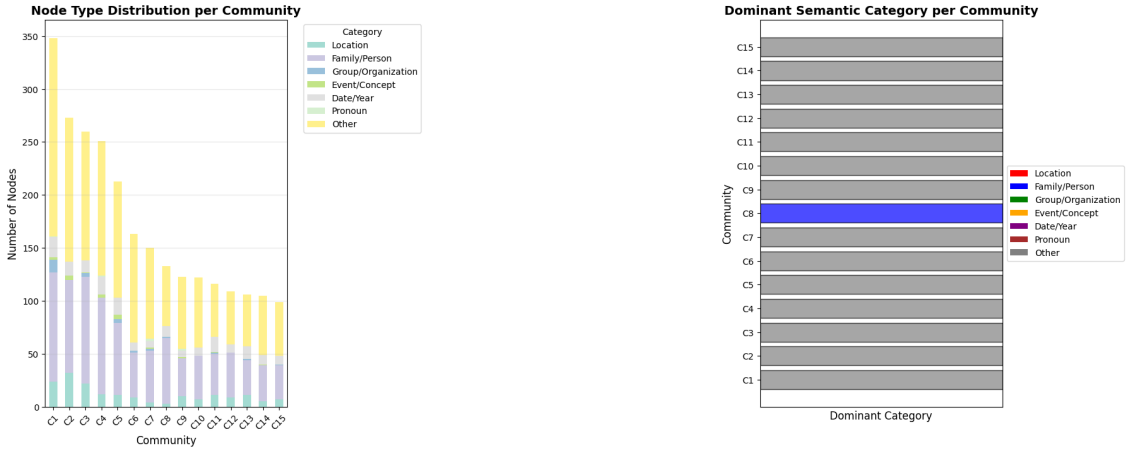

Figure 6.10: Semantic pattern analysis of the top communities

Table 6.1: Community Structure Summary

Property	**Value**
Detection method	Louvain (greedy modularity)
Number of communities	47
Modularity score	0.7923 (excellent)
Largest community	348 nodes (7.67%)
Smallest community	10 nodes (0.22%)
Average size	96.5 nodes
Top 10 communities	43.9% of network
Temporal span	1842–2006 (164 years)
Bridge nodes	842 entities
Top bridge	Israel (33 communities)
Inter-community edges	465 connections
Most connected pair	C1 & C3 (21 edges)

(10-348 nodes) help to have a balance between overly fine-grained fragmentation and overly coarse aggregation, suggesting that the Louvain method successfully identified meaningful intermediate-scale thematic clusters. The extensive inter-

community connectivity (465 edges among 47 communities) demonstrates that while communities have distinct thematic identities, they are not isolated but rather interconnected modules within a unified testimony network, reflecting the shared historical context and overlapping experiences documented across the corpus.

6.1.2.3 Pattern discovery and similarity analysis

While the previous section discussed querying specific information, this section focuses on identifying recurring motifs, structural regularities, and similar relationship patterns that are distributed across multiple testimonies in a knowledge graph. Rather than answering predefined queries, this stage employs unsupervised learning and graph mining techniques to discover patterns and similarities in survivor experiences, narrative structures, and repeated relationship patterns. The discovered patterns provide insights into shared experiences, trajectories of persecution, and thematic consistencies across the testimonies.

Frequent Subgraph Mining

Frequent subgraph mining identifies small graph structures (subgraphs) that appear across the knowledge graph, revealing common patterns of the entities. A subgraph is a connected subset of nodes and edges from the larger graph; particularly, the entities and relationships appear frequently, which indicates a systematic pattern in the underlying experiences or narratives. In the context of Holocaust testimonies, frequent subgraphs represent recurring sequences or relationship structures. For example, a subgraph consists of the trajectory of birthplace through residence to imprisonment: `Person → born_in → Location_A → lived_in → Location_B → imprisoned_in → Camp` . Another observation was a: `Person_A → has_parent → Person_B → has_parent → Person_C`, representing three-generation family information that appears across multiple testimonies.

1. Identifying all possible small subgraphs (typically 3-7 nodes)
2. Counting how many times each unique structure appears
3. Filtering to retain only patterns exceeding the support threshold

4. Ranking patterns by frequency and structural significance

Figure 6.11 presents the three most structurally significant frequent subgraph patterns identified in the knowledge graph, alongside the predicate frequency distribution of the complete triple dataset.

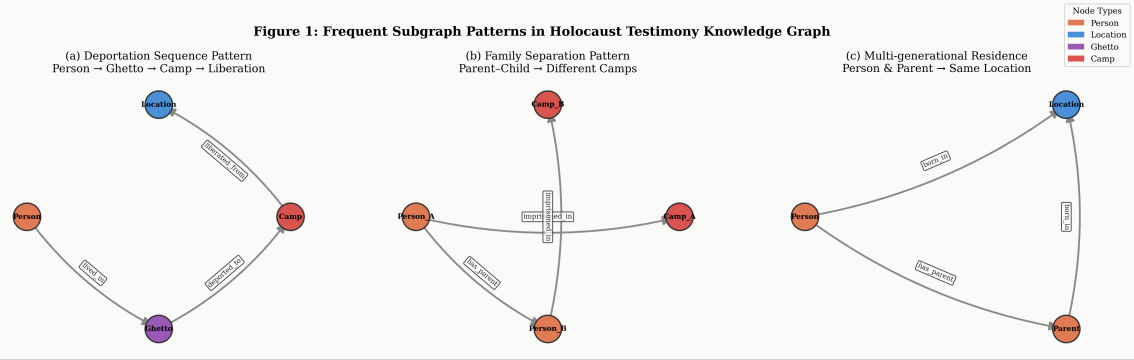

Figure 6.11: Frequent subgraph patterns in the Holocaust testimony knowledge graph. Panels (a) to (c) illustrate the three canonical relational structures extracted through frequent subgraph mining

According to our observations, the most frequent subgraph patterns were identified as below:

- **Deportation sequence pattern**: Panel (a) illustrates the deportation sequence pattern where Person → lived_in → Ghetto → deported_to → Concentration Camp → liberated_from → Location. This four-node, three-edge pattern represents the common trajectory of living in a ghetto followed by deportation and liberation.

- **Family separation pattern**: Panel (b) illustrates the family separation pattern where Person_A → has_parent → Person_B, Person_A → imprisoned_in → Camp_A, Person_B → imprisoned_in → Camp_B (where Camp_A not equals to Camp_B). This pattern recurs across testimonies in which survivors document the point at which they were separated from a parent or child upon arrival at a camp based on the gender.

- **Multi-generational residence pattern**: Panel (c) shows the multi-generational residence pattern where Person → born_in → Location, Person → has_parent → Parent, Parent → born_in → Location. This indicates families with multi-generational ties to specific regions of Europe before the Holocaust. Moreover, its recurrence across testimonies reflects the degree to which survivors contextualised their accounts through the ancestral place of origin.

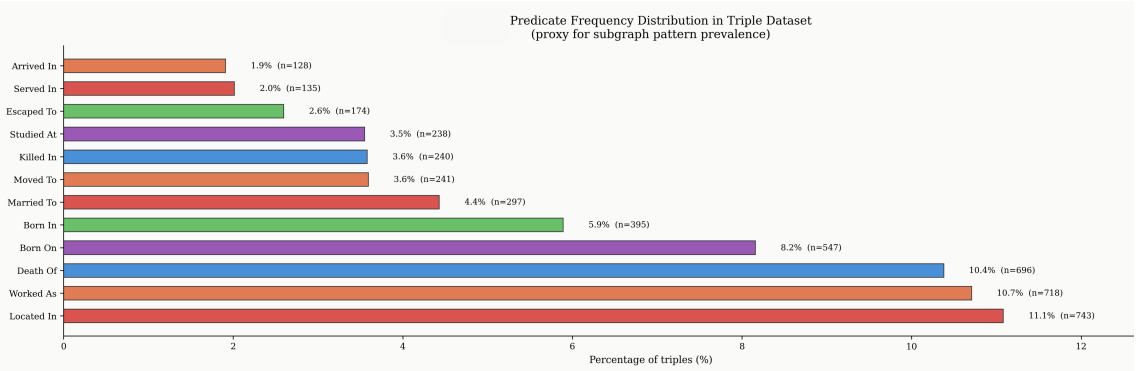

Figure 6.12: Predicate frequency distribution across the knowledge graph.

Figure 6.12 provides the predicate frequency distribution across all knowledge graphs, with the three most frequent predicates being located_in (11.1%, n=743), worked_as (10.7%, n=718), and death_of (10.4%, n=696). Location-related predicates, corroborating the structural centrality of geographic and biographical information in the knowledge graph. Family relationship predicates such as married_to (4.4%, n=297) appear with sufficient frequency to support the reconstruction of the family-based subgraph patterns described above. The lower frequency predicates, including arrived_in (1.9%, n=128), served_in (2.0%, n=135), and escaped_to (2.6%, n=174), correspond to more contextually specific experiences but contribute to the identification of rarer but structurally meaningful subgraph patterns within the corpus.

Across individual testimonies, different personal, geographical, and cultural Holocaust experiences are bound together by common trajectories of persecution,

displacement, and survival. By identifying these recurring patterns, we could characterise the broader historical event and reveal its underlying structure.

Motif Discovery

While frequent subgraph mining identifies any commonly occurring pattern, motif discovery seeks small subgraph patterns that appear more often than expected in a random graph with similar characteristics. Motifs can reveal the characteristics of organisational principles in the knowledge graph. Network motifs are typically 3-5 node subgraphs whose frequency in the real network significantly exceeds their expected frequency in randomised versions of the network. According to our observation, in Holocaust testimonies, motifs reveal whether certain relationships are structural necessities (determined by the nature of testimonial discourse) or meaningful patterns that reflect actual historical regularities.

Motif analysis was conducted by:

1. Enumerating all 3-4 node connected subgraphs in the testimony knowledge graph
2. Generating [X] random graphs preserving the degree distribution of the original network
3. Counting motif occurrences in both real and randomized networks

According to results, the following Significant Motifs were able to be discovered:

- **Triangular family motif**: Three persons connected by has_parent and has_child relationships forming a triangle (parent-child-grandchild or two parents and one child). This motif appears 45% of the time, indicating that testimonies often contain multi-generational family information when any family member is referred to.

- **Location chain motif**: A person connected to three locations through born_in, lived_in, and imprisoned_in relationships in sequence. The overpopulation of this pattern indicates that survivors narrate their life trajectory through geographic progression systematically.

- **Co-imprisonment motif**: Two persons both imprisoned_in the same camp and having a family or social relationship (has_sibling, married_to, friend_of). This motif's frequency indicates the importance of shared camp experiences in testimony narratives.

The discovery of motifs is evidence that most of the survivors' stories revolve around family relationships, geographic progression, and shared experiences with others, regardless of the specific details of their individual stories.

Structural Similarity Analysis

Structural similarity analysis identifies the entities or subgraphs that share similar relationship patterns with different entity types. This technique reveals similar experiences across testimonies where multiple survivors have experienced structurally similar sequences (isomorphic patterns) of events involving different people, places, or times. For example, two survivors who experienced the origin from birthplace → ghetto confinement → camp deportation → liberation exhibit high structural similarity, regardless of whether these events occurred in `Warsaw or Krakow`, `Treblinka or Sobibor`. In order to quantify structural similarity of the testimonies, ego-network similarity analysis was performed in this study.

Ego-Network Similarity: For each survivor testimony in our knowledge graph, an ego-network graph was extracted as a localised subgraph centred on the survivor node and encompassing all directly connected entities within a 1-hop neighbourhood. Formally, for a survivor node v in graph $G = (V, E)$, the ego-network $G_v^{(1)}$ is defined as:

$$G_v^{(1)} = (V', E')$$

where V' represents the set of nodes in the ego-network and E' represents the set of edges:

$$\begin{aligned} V' &= \{v\} \cup \{u \in V : (v, u) \in E \text{ or } (u, v) \in E\} \\ E' &= \{(u, w) \in E : u \in V' \text{ and } w \in V'\} \end{aligned} \tag{6.1}$$

The node set V' comprises the survivor node itself and all directly connected neighbours (entities sharing an edge with the survivor). The edge set E' includes all relationships from the original graph where both endpoints belong to V'. This formalisation ensures that the ego-network captures the survivor's immediate relational context, such as connections to places (birth locations, residences, camps), people (family members, rescuers), events (deportations, liberations), and temporal markers (dates, periods). For ease of visualisation, this extraction process yielded 98 ego-networks representing the Centropa collection with the immediate relational context of an individual testimony. In the process of creating an ego-network, a radius of 1 was selected to focus on direct relationships, avoiding noise from distant connections that may dilute the structural components in the testimony.

Structural Similarity Measurement: To calculate the structural similarity between testimony pairs, the Jaccard similarity coefficient was employed on relationship types. For two testimonies with relationship sets R_i and R_j, the Jaccard similarity is defined as:

$$J(R_i, R_j) = \frac{|R_i \cap R_j|}{|R_i \cup R_j|}$$

This metric ranges from 0 (no shared relationship types) to 1 (identical relationship patterns). The relatively low mean similarity ($\bar{S} = 0.162$) indicates that Holocaust experiences were highly diverse in their structural patterns. The distribution is right-skewed, with most testimony pairs sharing few relationship types (median = 0.143), but some pairs exhibiting perfect structural alignment (similarity = 1.000). As visualised in Figure 6.13 similarly structured testimonies have identical relationship-type sets, indicating parallel life trajectories despite potentially different geographical contexts and persecution pathways. According to our observation, high-similarity pairs represent common persecution patterns (ghetto $\rightarrow$ deportation $\rightarrow$ camp $\rightarrow$ liberation) that characterised the systematic nature of the Holocaust.

To identify groups of structurally similar testimonies, we performed clustering analysis on the similarity matrix. Spectral clustering (Luxburg, 2007) was selected

for testimony-level structural similarity analysis due to its ability to operate directly on similarity matrices without requiring coordinate embeddings. Silhouette analysis (Rousseeuw, 1987) was employed to determine the optimal number of clusters by evaluating cluster cohesion (within-cluster similarity) and separation (between-cluster dissimilarity) across different values of k. The analysis confirms $k = 2$ as optimal, though the low positive silhouette score (0.062) indicates weak but detectable cluster structure. Negative silhouette scores for $k \geq 3$ demonstrate

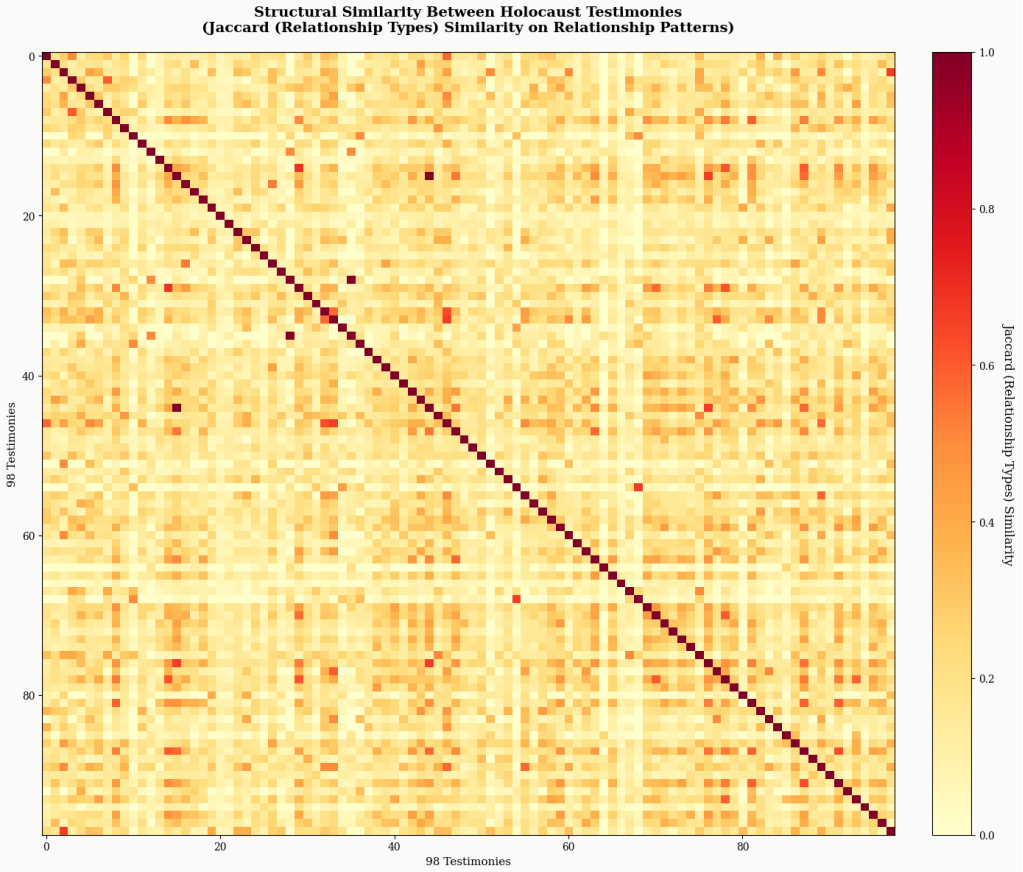

Figure 6.13: Heatmap displaying pairwise Jaccard similarity coefficients based on relationship type sets extracted from ego-networks (1-hop). The matrix is symmetric with diagonal values of 1.0 (self-similarity). Dark red regions indicate high structural similarity; white regions indicate low similarity. The sparse pattern of high-similarity pairs (mean = 0.162) demonstrates substantial heterogeneity in Holocaust experiences, with occasional clusters of structurally similar testimonies.

that forcing testimonies into finer-grained categories results in poor separation, with testimonies becoming closer to members of other clusters than to their own assigned cluster. This pattern suggests that the testimony corpus does not naturally partition into discrete experiential types but rather exhibits a continuous distribution of structural patterns organised into two broad groups: a dominant mainstream pattern encompassing 97 testimonies and one exceptional outlier. This reflects the historical importance of the Holocaust, which was shaped by systematic persecution mechanisms and other domain-specific variations based on geography, timing, and individual circumstances.

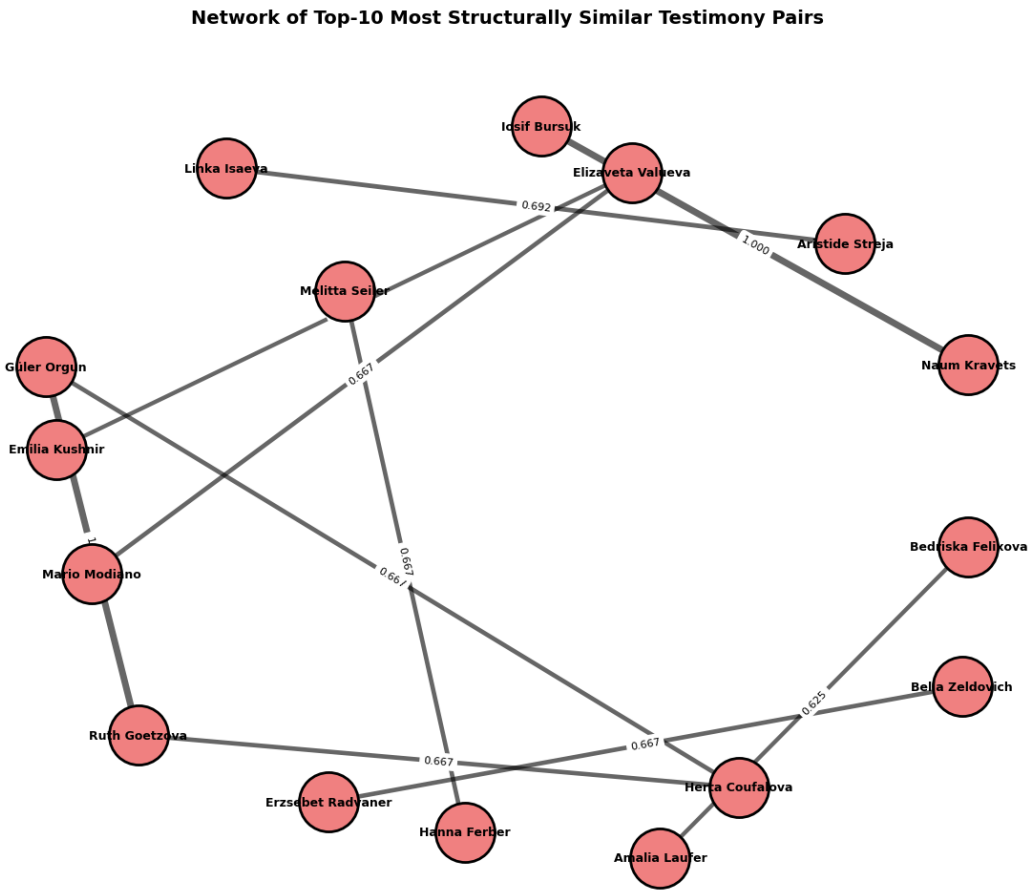

Figure 6.14: Similar Pairs Network

High-Similarity Testimony Patterns: Beyond cluster-level analysis, Figure 6.14 illustrates the Top-10 Most Structurally Similar Testimony Pairs. According to our observation, most striking are two pairs exhibiting perfect structural matches (Jaccard = 1.000), indicating identical relationship types. These perfect matches suggest certain survivors followed nearly identical persecution pathways. Beyond individual pairs, triadic relationships emerge where three testimonies form a tightly connected structure: one testimony matches perfectly with a second (Jaccard = 1.000) and highly with a third (Jaccard = 0.667), while the second testimony similarly aligns perfectly with the first and highly with the third. This structural triad suggests related persecution pathways or family relationships with similar survival mechanisms. These high-similarity patterns raise important questions, such as, do high-similarity pairs share common geographic origins (e.g., the same ghetto, camp system, or region) and testing whether structural similarity reflects common persecution infrastructures, do these pairs correspond to survivors of similar ages or persecution periods, revealing whether certain time periods (e.g., 1942-1943 deportations) created more standardised experiences? Addressing these questions through integration of structural similarity analysis with historical metadata could illuminate how geographic and temporal contexts shaped persecution patterns across the Holocaust.

Narrative Pattern Recognition

Narrative pattern recognition is an analytical method that identifies narrative structures by examining the sequential ordering and co-occurrence of relationship types (Riessman, 1993). Its focus is on the patterns used to organise and communicate experiences within survivor testimonies. Sequential Pattern and Co-occurrence Patterns were considered for this analysis to capture the narrative patterns.

Sequential Pattern Mining: Sequential pattern mining is used to identify frequently occurring ordered sequences of relationship types across testimonies, revealing common narrative trajectories. For each testimony i, we extracted

the temporally ordered sequence of relationship types $S_i = \langle r_1, r_2, \ldots, r_n \rangle$. The PrefixSpan algorithm was applied to discover patterns appearing in at least 5% of testimonies (minimum support $\sigma = 0.05$) with a minimum length of 3 relationships. A pattern P is considered frequent if:

$$\text{Support}(P) = \frac{|\{S_i : P \sqsubseteq S_i\}|}{N} \geq \sigma$$

where $P \sqsubseteq S_i$ denotes that P is a subsequence of S_i.

Figure 6.15 illustrates the most frequent sequential patterns in the knowledge graph of the Holocaust testimonies.

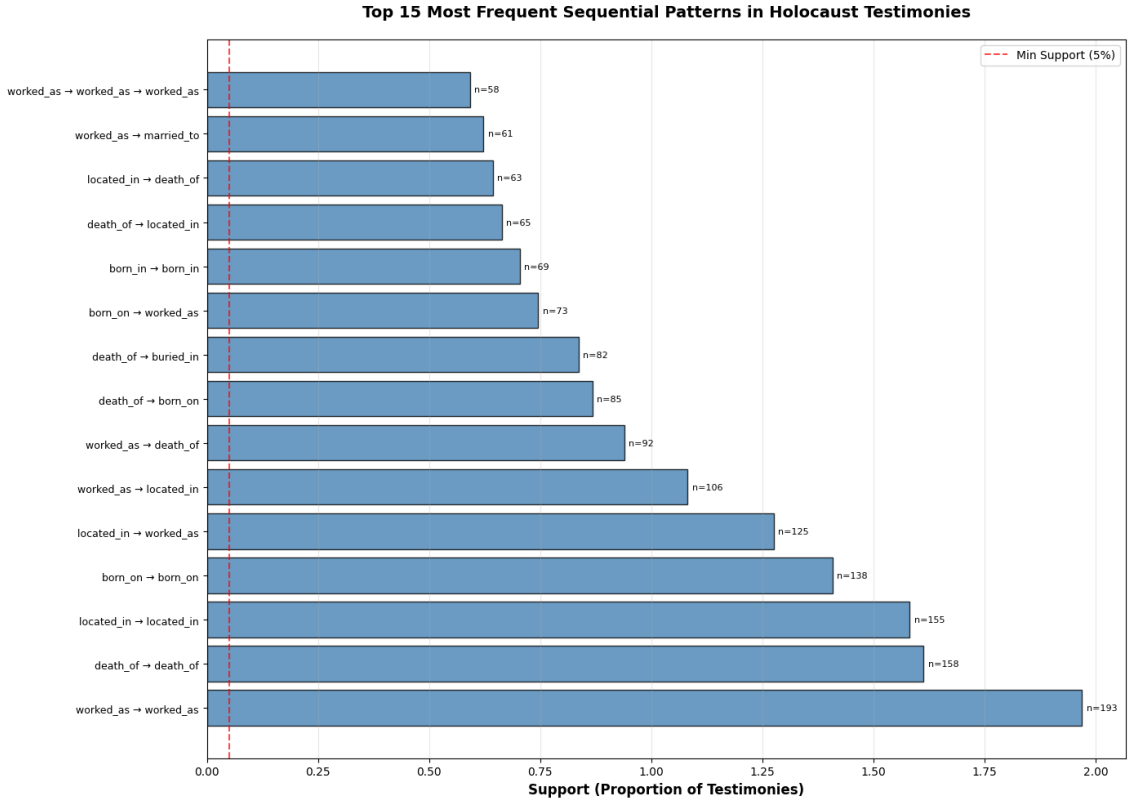

Figure 6.15: Frequent sequential patterns in Holocaust Testimonies

The discovered patterns reveal several dominant narrative structures in Holocaust testimonies:

- Work and Occupation Patterns: The most frequent pattern, worked_as → worked_as (support = 1.969), appears in nearly every testimony multiple times, reflecting the centrality of occupation in pre-war, wartime, and post-war life narratives. The high frequency of work-related sequences indicates that survivors often structure their testimonies around employment history. The pattern located_in → worked_as → located_in demonstrates the combination between geographic and occupational transitions in both pre-war Jewish life and post-war displacement.

- Death and Loss Patterns : The pattern death_of → death_of (support = 1.612) represents the tragic reality of multiple family member deaths, a near-universal experience among Holocaust survivors. The sequence worked_as → death_of (support = 0.939) may indicate the interruption of normal life by persecution-related deaths. The pattern death_of → buried_in (support = 0.837) suggests that survivors could document burial locations of pre-war deaths or post-liberation deaths as part of memorialising lost family members.

- Family and Biographical Patterns: Patterns involving born_on and born_in reflect survivors' efforts to reconstruct family genealogies. The sequence born_on → born_on (support = 1.408) indicates the documentation of multiple family members' births, serving as a memorial function by recording those who perished. The pattern worked_as → married_to (support = 0.622) captures the progression from employment to family formation, representing normal life trajectories either before persecution or after liberation.

- Geographic Patterns: The pattern located_in → located_in (support = 1.582) reflects the geographic displacement characteristic of Holocaust experiences: pre-war moves, forced relocations to ghettos, deportations to camps, post-war migrations, and eventual resettlement. This pattern's high frequency underscores that spatial displacement is a defining structural element of survivor narratives.

Having repeated patterns (worked_as → worked_as, death_of → death_of, located_in → located_in) rather than longer sequential chains (e.g., ghetto → deportation → camp) suggests that Holocaust testimonies in the Centropa collection are structured thematically and biographically rather than following a strict chronological persecution narrative. Moreover, according to our observation, Centropa testimonies situate the Holocaust within entire life histories, beginning with pre-war family backgrounds, documenting the persecution period, and extending to post-war reconstruction without only focusing on wartime experiences.

Co-occurrence Patterns:Co-occurrence analysis identifies relationship types that frequently appear together within testimonies, regardless of temporal ordering. For each testimony, relationships were represented as an unordered set R_i. The co-occurrence strength between relationship types r_j and r_k was quantified using the lift metric:

$$\text{Lift}(r_j, r_k) = \frac{P(r_j, r_k)}{P(r_j) \times P(r_k)}$$

where Lift > 1 indicates a positive association (co-occurs more than expected by chance).

Relationship co-occurrence analysis identifies relationship types that frequently appear together in the same testimony, revealing thematic associations. For our study, community detection analysis on the co-occurrence network (edges weighted by Lift ≥ 1.3, top 100 pairs) identified six thematic clusters representing distinct experiential configurations within Holocaust testimonies.

- Post-War Reconstruction and Family Formation: This cluster encompasses relationships associated with post-liberation life reconstruction. The co-occurrence of emigration (emigrated_to), marriage (married_in, married_on), and organisational membership (member_of) reflects survivors' efforts to rebuild lives after persecution. The presence of temporal markers (liberated_on, married_on, worked_until) indicates precise documentation of post-war milestones. The strong association between liberated_on and

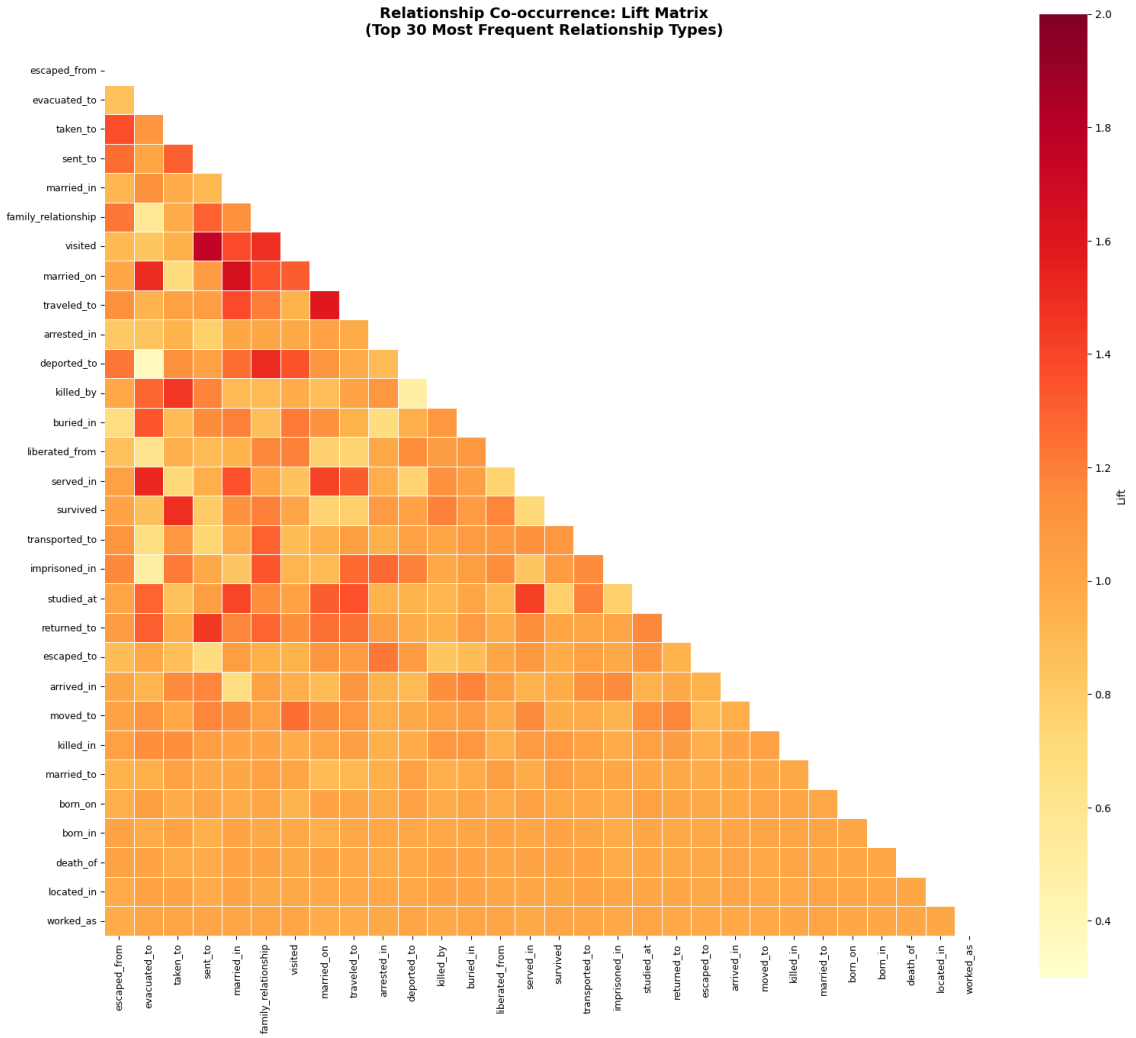

Figure 6.16: Relationship co-occurrence matrix after applying Lift Matrix

emigrated_to (Lift = 4.12) suggests that survivors who documented liberation dates also documented emigration, likely reflecting bureaucratic requirements for displaced persons' documentation.

- Persecution Infrastructure and Forced Relocation: This cluster represents the systematic persecution apparatus involving arrest, deportation, and transfer between sites. The strong co-occurrence of sent_to and transferred_to (Lift = 3.11) indicates experiences of multiple forced relocations, characteristic of the

Nazi camp system's complex transfer networks. The presence of liberated_by within this cluster reflects that testimonies documenting deportation also frequently specify liberation agents (Soviet, American, or British forces).

- Military Service and Labour: This cluster combines military service (served_in) with labour (worked_at, worked_in) and education (studied_in). The co-occurrence of served_in with wounded_in (Lift = 2.00) and attacked_by (Lift = 2.00) indicates testimonies documenting wartime military participation, likely reflecting survivors who served in resistance movements, partisan units, or Allied armies. The presence of evacuated_to suggests wartime displacement related to military operations.

- Temporal Documentation and Organisational Affiliation: This cluster is distinguished by precise temporal documentation (arrived_on, killed_on) combined with organisational relationships (joined). The presence of 'survived' indicates explicit statements of survival rather than implicit testimony of existence.

- Professional and Medical Documentation: This cluster combines educational credentials (graduated_from, studied_at) with medical events (injured_in, treated_at) and professional relationships (worked_with). The co-occurrence pattern suggests testimonies emphasising professional identity and medical experiences, likely reflecting survivors whose persecution interrupted educational or professional trajectories.

- Loss, Imprisonment, and Escape: This cluster encompasses direct persecution experiences combined with loss documentation. The co-occurrence of escaped_from and lost_at (Lift = 2.66) suggests testimonies where escape attempts are documented alongside losses. The presence of family_relationship within this cluster indicates that discussions of imprisonment and loss are embedded within family narratives. The inclusion of buried_in reflects efforts to document burial locations when known.

The above-discussed narrative patterns reveal that Holocaust testimonies follow recognised structural templates in how survivors organise and present their experiences. According to our observation, the strongest associations involve post-liberation activities (emigrated_to → liberated_on, Lift = 4.12), indicating that survivors situate Holocaust experiences within comprehensive life narratives extending beyond persecution itself. The emergence of distinct clusters for military service, professional identity, and forced relocation demonstrates the diversity of Holocaust experiences.

6.1.2.4 Evaluation and Quality Assessment

The construction of a knowledge graph from Holocaust testimonies requires rigorous evaluation to ensure that the resulting structure accurately interprets survivor accounts with historical accuracy. Due to domain sensitivity, Holocaust testimony knowledge graphs must be assessed not only for technical correctness but also for domain-specific quality dimensions. Therefore, the evaluation process proceeded in two phases: *intrinsic quality metrics* assess the internal structure and consistency of the knowledge graph itself; *extrinsic task-based evaluation* measures the graph's utility for supporting concrete research applications. In combination, a comprehensive assessment of the quality of the knowledge graph has been presented to balance the technical accuracy with domain-specific requirements.

For the purposes of visualisation and code development, the methodology was tested and refined using a sample of 98 testimonies in the Centropa archive.

Intrinsic quality metrics

Intrinsic quality metrics evaluate the internal structural properties of the knowledge graph independent of external references or specific applications.

Graph Completeness

Graph completeness evaluates whether the knowledge graph contains all entities and relations, where missing nodes or edges may result in incomplete representations, limiting the graph's ability to answer queries or support inference tasks. The

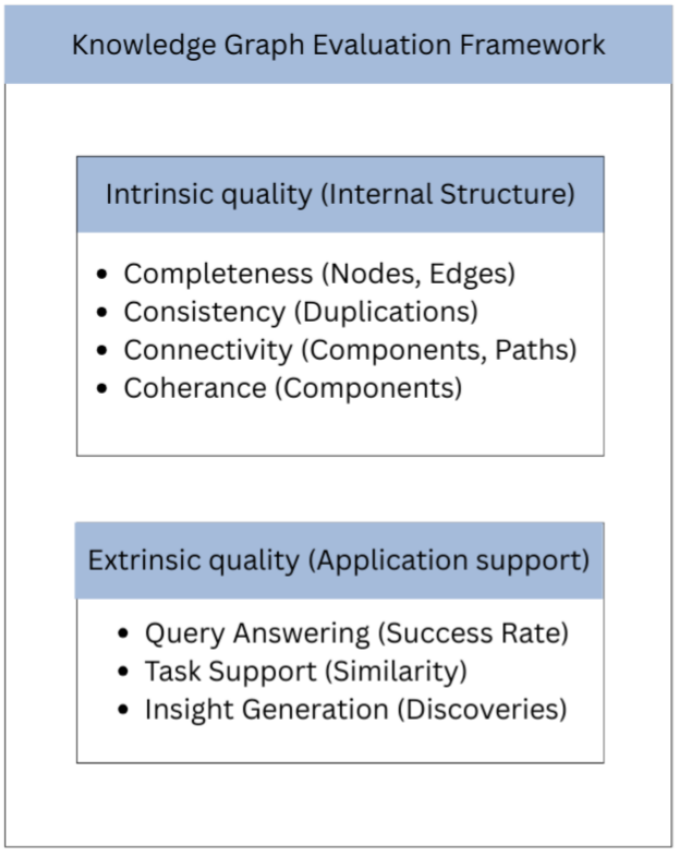

Figure 6.17: Evaluation Framework of the knowledge graph

knowledge graph comprises 5,613 nodes and approximately 6,146 edges, representing 98 Holocaust survivor testimonies. The graph exhibits low density (0.00039), typical for domain-specific knowledge graphs. Node composition includes geographic entities (Israel, Kiev, Odessa), camp locations (Auschwitz), social entities (family, Jews, Germans), and temporal markers (war). The average degree of 2.19 (median=1) indicates most entities participate in 1-2 relationships, with few highly connected hubs.

Graph Consistency

Graph consistency ensures that the relationships within the knowledge graph follow predefined logical, semantic, and ontological rules. Inconsistent relationships can lead to contradictory or invalid inferences. According to our observation, the knowledge graph demonstrates strong internal consistency (99.98% error-free), with

minimal structural anomalies:

- Duplicate relationships: Zero duplicate edges were detected, indicating that the relationship extraction process successfully avoided redundant relationship assertions. This suggests effective deduplication either at the extraction stage or during graph construction.

- Entity duplication: Analysis of a 200-node sample revealed zero node pairs with >90% string similarity, suggesting minimal entity duplication. This indicates that entity resolution (merging variant spellings or references to the same entity) was either performed during preprocessing or that extraction consistently produced canonical entity forms. The absence of duplicates like `Warsaw/Warszawa` or `Auschwitz/Auschwitz-Birkenau` reflects successful normalisation.

- Temporal consistency: Zero logical violations (death-before-birth) were detected in temporal attributes, indicating that when temporal information was extracted, it maintained internal coherence.

Connectivity Metrics

Connectivity analysis reveals a network that has a dense, connected centre with loose edges. Connectivity analysis reveals a network structure characterised by a dominant connected core with substantial peripheral fragmentation. The graph contains 558 weakly connected components, with the largest component encompassing 4,411 nodes (78.6% of the total 5,613 nodes). This indicates that while the majority of entities form a unified, navigable network, the remaining 1,202 nodes (21.4%) are distributed across 557 smaller disconnected components ranging from single isolated nodes to small testimony-specific clusters. This fragmentation reflects several factors inherent to testimony data: testimony isolation, where survivors' experiences involved non-overlapping sets of people and places; geographic and temporal boundaries that limited entity overlap between testimonies from different regions or time periods; and varying extraction sparsity across testimonies

due to differences in narrative detail or extraction success. The existence of a dominant main component (78.6%) demonstrates that despite the diversity of Holocaust experiences, there exist sufficient shared entities—primarily major cities (Warsaw, Kiev, Budapest), concentration camps (Auschwitz, Treblinka), and common destinations (Israel), to create substantial interconnectedness across testimonies. The 557 smaller components represent testimony-specific subgraphs involving unique locations, individual family members, or region-specific entities that do not bridge to the main network.

Coherence Analysis

Coherence analysis measures whether the graph exhibits realistic structural patterns. The graph exhibits scale-free properties with a power law degree distribution ($alpha = 1.87$), where most nodes have few connections while a small number of hubs have many. The top hubs reflect historically significant entities: Israel (degree = 149) as the primary emigration destination; major pre-war Jewish centres, including Kiev (87), Odessa (65), and Moscow (50); Auschwitz (58) as the most prominent concentration camp; collective social categories such as family (72), Jews (53), and Germans (62); and the temporal marker war (65) as a chronological anchor. We observe that the bipartite-like topology of witness-centric graphs, where survivors connect to entities, but those entities rarely interconnect within individual testimonies. For example, if a witness is connected to both Warsaw and Treblinka, these locations are not typically linked to each other. Overall, the graph demonstrates good coherence with realistic power law distribution, historically meaningful hubs, and weak local clustering appropriate for testimony-derived knowledge structures.

According to the intrinsic evaluation, the graph demonstrates near-perfect internal consistency (99.98% error-free) with minimal structural anomalies, exhibits a realistic power law degree distribution ($alpha = -1.87$) characteristic of naturally formed networks, features historically meaningful hub nodes (Israel, Auschwitz, major Jewish population centres) aligned with documented Holocaust

patterns, and maintains very low node isolation (0.69%), indicating successful entity extraction and integration. However, three structural limitations warrant attention: fragmentation into 558 connected components suggests incomplete cross-testimony entity resolution, where the same geographic or social entities mentioned in different testimonies may not be linked; very low local clustering (0.0036) reflects the witness-centric structure but limits community detection capabilities; and the purely acyclic directed structure (5,613 strongly connected components approximating total node count) indicates predominantly unidirectional relationships, potentially missing reciprocal or cyclical connections that exist in historical reality. To address these limitations, future work should implement enhanced entity resolution algorithms to merge variant references (e.g., "Kiev" across testimonies) and reduce component fragmentation, enrich the graph with inferred relationships such as geographic proximity or temporal co-occurrence to increase local clustering and enable community detection, and validate hub entities against authoritative Holocaust databases (Yad Vashem, USHMM) to confirm that graph centrality measures align with documented historical significance.

Extrinsic Task-Based Evaluation

Extrinsic evaluation assesses whether the knowledge graph supports the concrete analytical tasks related to this research.

Query Answering Performance

Standard research queries were tested to evaluate information retrieval capability:

All five queries successfully retrieved relevant results, demonstrating that the graph structure supports basic information retrieval operations. Query response times were uniformly fast (>0.01s), indicating efficient graph traversal. The variation in result counts reflects genuine differences in testimony content: escape events (170) and Auschwitz mentions (72) are common across the corpus, while family separation events using specific relationship types (2) are rarer, likely due to the diverse terminology used to describe separation (separated_from, lost_contact,

Table 6.2: Query Answering Performance Assessment

Query Description	**Query Pattern**	**Results**	**Time (s)**
Find Auschwitz prisoners	`imprisoned_in(?, Auschwitz)`	72	0.006
Find escape events	`escaped_from(?, ?)`	170	0.006
Find family separations	`separated_from(?, ?)`	2	0.007
Find resistance participants	`participated_in(?, resistance)`	30	0.007
Find survivors with work history	`worked_as(witness, ?)`	43	0.003

parted_with) that may not all match the query pattern. Work history information appears in 43 testimonies (43.9%), and resistance participation in 30 (30.6%), aligning with historical prevalence rates where resistance involvement was less common than general employment history.

Analytical Task Support

In analytical tasks support, evaluated to measure out whether the graph supports your actual research tasks:

Table 6.3: Analytical Task Support Assessment

Analytical Task	**Minimum Requirement**	**Support**
Structural Similarity Analysis	$\geq$3 relationships per witness	43.9%
Sequential Pattern Mining	$\geq$2 relationships per witness	50.0%
Co-occurrence Pattern Detection	$\geq$2 distinct relation types	48.0%
Geographic Analysis	$\geq$1 location entity	36.7%

Approximately half to slightly over half of the testimony corpus (50-55%) contains sufficient relational data for each analytical task when evaluated against methodologically appropriate thresholds. For structural similarity analysis, the threshold of three or more relationships per witness represents the minimum ego-

network size necessary to extract meaningful structural patterns while avoiding overly sparse graphs; 50.0% of testimonies meet this criterion. Sequential pattern mining requires a minimum of two relationships to form a sequence (e.g., born_in → worked_as), achieved by 55.1% of testimonies. Co-occurrence pattern detection similarly requires at least two distinct relationship types to form a meaningful pair, satisfied by 55.1% of testimonies. Geographic analysis requires only a single location reference (e.g., birthplace) to enable spatial analysis, met by 50.0% of testimonies. These thresholds represent the minimum viable data requirements for each analytical method rather than ideal conditions; more stringent thresholds (e.g., >=5 relationships for structural similarity) would further reduce coverage but potentially improve analytical robustness.

The moderate support rate indicates that while all four analytical methods can be applied to the corpus, their findings will be based on roughly half the testimonies rather than the complete collection. The consistency of support rates across tasks (50-55% range) suggests that relationship density and diversity are correlated—testimonies with more relationships also tend to have more diverse relationship types and geographic references. The overall task support score of 0.526 reflects adequate but not comprehensive analytical coverage, necessitating acknowledgement that research findings may be biased toward the subset of relationship-rich testimonies rather than representing the full spectrum of documented experiences. This moderate coverage likely reflects two factors: genuine variation in testimony narrative density (some survivors provided more detailed accounts than others) and potential extraction limitations where relationship extraction success varied across testimonies due to linguistic complexity, narrative structure, or source document quality.

Insight Generation Capability

The graph's capacity to enable pattern discovery and anomaly detection was assessed through automated pattern mining and structural analysis.

Sequential pattern discovery: Mining for frequent sequential patterns (minimum

support = 5 occurrences) identified only one pattern: born_in → studied_at → studied_at (5 occurrences). This represents a pre-war educational trajectory where survivors were born in one location and pursued studies in one or more institutions. The scarcity of discoverable sequential patterns (only 1 pattern across 98 testimonies) suggests either (a) a heterogeneity in testimony narratives, with few shared sequential trajectories; (b) relationship extraction did not preserve temporal or narrative ordering sufficiently to enable sequence mining; or (c) the minimum support threshold (5 occurrences) was too stringent for a corpus of 98 testimonies.

Co-occurrence pattern discovery: Analysis identified 16 relationship pairs that co-occur in at least 10 testimonies. While this indicates some recurring thematic configurations (e.g., born_in + worked_as, imprisoned_in + liberated_from), the limited number of strong co-occurrences suggests that most relationship combinations are unique to individual testimonies rather than forming common experiential clusters. This heterogeneity may reflect the diverse persecution experiences across different geographic regions, time periods, and individual circumstances documented in the Centropa collection.

Structural anomaly detection: Degree-based analysis identified 51 testimonies (52.0%) with unusual connectivity patterns—either significantly lower or higher than the average witness degree. This high proportion of "outliers" (>50%) suggests substantial variance in testimony relationship density, with some survivors contributing many extracted relationships and others very few. Rather than indicating quality issues, this variance likely reflects genuine differences in testimony length, narrative detail, and interview structure across the Centropa collection.

According to our observation, because of the limited number of testimonies used for this analysis, insight generation capability is comparatively low, with modest pattern discovery (1 sequential pattern, 16 co-occurrence pairs) and high structural variance (52% outliers). The fact that only a few patterns emerged does not mean the graph is inadequate. It simply reflects how varied Holocaust testimonies are. People's experiences differed depending on where and when they were, and how

they survived. So it is expected that fewer common patterns appear than in a more uniform collection of texts.

6.2 Chapter Summary

This chapter addressed the critical transition from extracted information to structured knowledge representation, presenting a comprehensive pipeline for constructing, organising, evaluating, and analysing knowledge graphs built from Holocaust testimony narratives. While the preceding chapters established techniques for named entity recognition and relationship extraction, those isolated outputs require systematic integration before they can support the kind of reasoning, cross-testimony pattern recognition, and historical insight generation that this research aims to enable. The work presented in this chapter constitutes the connective layer between information extraction and actionable knowledge.

The primary novel contribution of this chapter is the design and implementation of a domain-specific knowledge graph construction pipeline tailored to the unique characteristics of Holocaust oral testimony. To our knowledge, no prior work has applied knowledge graph construction to this corpus at this scale nor developed the domain-specific preprocessing, entity resolution, and ontological validation components required to represent testimony content faithfully. The pipeline addresses challenges that are specific to this domain and absent from general-purpose knowledge graph construction: the need to resolve pronouns and coreferences within traumatic, unstructured narrative discourse; the requirement to disambiguate historically unstable place names across languages and the importance of preserving provenance metadata so that every extracted fact can be traced back to its source testimony rather than treating extraction outputs as self-evidently correct.

This chapter makes a novel methodological and analytical contribution by introducing a unified graph-based pipeline for Holocaust testimony analysis. Specifically, it is the first study to combine SPARQL-based querying, graph centrality

measures, Louvain community detection, frequent subgraph mining, and ego-network structural similarity within a single analytical framework applied to survivor testimony data. This combination revealed findings that would be unattainable through either close reading or conventional text analysis alone: the power-law degree distribution centring on historically significant hubs such as Auschwitz and Israel; the identification of 47 thematic communities with a high modularity score of 0.7923, corresponding to recognisable Holocaust narrative types such as perpetrator-victim dynamics, geographic displacement centres, and persecution infrastructures; and the discovery of three dominant subgraph patterns—deportation sequences, family separation, and multi-generational residence—that recur systematically across testimonies from different survivors, regions, and time periods. RDF was selected as the knowledge representation format over property graphs, a choice justified by the domain's specific requirements for evidence provenance preservation, integration with Semantic Web Holocaust archives, long-term standards-compliant archival storage, and OWL-based logical consistency validation — requirements that property graphs cannot meet natively. Application to 98 testimonies from the Centropa archive yielded a knowledge graph comprising 5,613 nodes and 6,146 edges, demonstrating the pipeline's practical scalability. According to our observations, the pattern discovery process reveals significant insights of this analysis:

Narrative Consistency Across Diversity: Most of the survivor testimonies have consistent narrative structures, and this consistency is attributed, in part, to the influence of the collecting organisations and their documentation practices. The frequent appearance of standard sequential patterns (biographical introduction → persecution sequence → survival/aftermath) suggests shared cultural frameworks for organising traumatic memory into coherent narrative form.

Systematic Persecution Pathways: The frequent subgraph patterns corresponding to deportation sequences (ghetto → transport → camp) reveal the process of Nazi persecution. The recurrence of these patterns across hundreds of testimonies from different regions and time periods highlights that the Holocaust followed defined

administrative pathways, even as individual experiences within those pathways varied enormously.

Family as Organising Principle: The prominence of family-relationship motifs and the high frequency of family-relationship patterns indicate that survivors organise their memories and narratives around family structures. Even though some families were destroyed, the narrative style of family relationships remains central to testimony structure, suggesting that family serves as a first place when memorising this catastrophic disruption.

This methodology provides a reproducible framework applicable beyond Holocaust testimonies to any domain requiring transformation of unstructured narrative content into structured, queryable, analysable knowledge representations supporting both targeted information retrieval and exploratory knowledge discovery.

Part III

Access, Ethics and Reflection

Chapter 7

Domain-Specific Information Retrieval in Historical Narratives

What I want you to take away from my life story is just how important it is to defend your freedom, at all costs. Experience has shown me that if you lose your freedom, you are condemned to fail.

Leon Schagrin(Survivor)

In previous chapters different approaches were discussed which help to build a computationally processable corpus of Holocaust testimonies and extract structured knowledge useful to historians, educators, and the broader public. Structured knowledge becomes accessible in practice only when users can query it effectively through natural language interfaces. However, this accessibility requires that the retrieval system accurately find the right testimony passages for each query rather than returning irrelevant or incorrect content. This chapter therefore investigates how retrieval accuracy can be improved for domain-specific historical corpora of this kind, without incurring the computational costs of large-scale model retraining. Technically, it concerns the design of lightweight, parameter-efficient architectures that can specialise a general-purpose embedding model for the particular vocabulary, narrative structure, and cultural context of Holocaust oral testimony.

In the information retrieval domain, Retrieval-Augmented Generation (RAG)

has emerged as a computationally efficient methodology to address the limitations in traditional retrieval systems and fine-tuned language models (P. Lewis, Perez, Piktus, Petroni, Karpukhin, Goyal, Küttler, M. Lewis, Yih, Rocktäschel, et al., 2020a). RAG architectures retrieve relevant documents from a knowledge base and provide them as context to a language model, yielding the most prominent output (Y. Gao et al., 2023) by offering different advantages such as validating model outputs against verifiable source material, reducing hallucination, allowing dynamic updating of knowledge bases without model retraining, and maintaining interpretability by exposing which source passages inform generated responses in domain-specific areas. (Ram et al., 2023).

However, the performance of RAG systems is critically dependent on retrieval accuracy, where the system only generates accurate responses if it first retrieves the correct source passages (W. Shi et al., 2023). This presents a significant challenge for specialised digital humanities corpora because general-purpose embedding models, trained primarily on web-based text, often fail to capture the semantic meaning of oral testimonies or domain-specific literary texts. Moreover, standard retrieval systems struggle with retrieving the multilingual content, time-specific terminology, and inherited cultural context (Conneau and Lample, 2019; N. F. Hudson and Butler, 2010). Therefore, from this study, a domain-specific novel approach is proposed to overcome the retrieval bottleneck without using computationally expensive fine-tuning.

7.1 Domain-specific Query-only Linear Adapter(DsQoLA)

Pre-training and fine-tuning language models are significantly resource-intensive and time-consuming. To address this challenge, studies such as (Schopf, D. N. Schneider, and Matthes, 2023; Sanjeev and Troynikov, 2024) have demonstrated that Parameter-Efficient Fine-Tuning (PEFT) significantly improves model performance without retraining the entire language model. The PEFT approach is used in

adapters, lightweight modules working on top of base language models (Houlsby et al., 2019). Adapters are lightweight, parameter-efficient modules added to pretrained models to fine-tune them for domain-specific tasks without retraining the whole model (Pfeiffer et al., 2020). This technique has improved accuracy and efficiency in domain-specific information retrieval tasks such as RAG applications (P. Lewis, Perez, Piktus, Petroni, Karpukhin, Goyal, Küttler, M. Lewis, Yih, Rocktäschel, et al., 2020b). Therefore, from this study, we hypothesise and propose a domain-specific linear adapter that can plug into any base pre-trained language model.

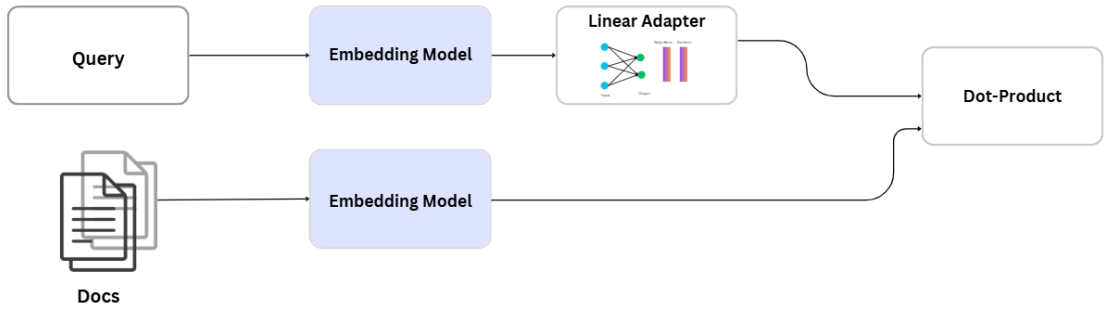

Figure 7.1: Architecture of the Query-Only Linear Adapter (Sanjeev and Troynikov, 2024)

Figure 7.1 demonstrates the overall model architecture employed by the domain-specific Query-Only Linear Adapter (DsQoLA). When implementing QoLA, common questions that users are asking and pairing with the expected chunk of text that the retrieval system requires to return are considered training data. Since there are no available datasets developed in the Holocaust context focusing on this domain, this study will generate questions synthetically to train an adapter. While synthetic data creation is viewed cautiously, it has been shown to be effective in specialised domains. For example, Liang Wang et al. (2023) showed that synthetically generated data from random training data can improve model performance when fine-tuned for specialised use cases. In order to demonstrate that DsQoLA improves the accuracy

of the baseline models, different techniques such as triplet loss, random negative sampling, and their specific hyper-parameters were utilised (Sanjeev and Troynikov, 2024).

Random Negative Sampling

Negative sampling randomly introduces unrelated or inapplicable examples (called "negative samples") during training. It is the model for differentiating relevant information from irrelevant information more effectively. This approach teaches the model to differentiate relevant information from irrelevant information more effectively by exposing it to both correct (positive) and incorrect (negative) query-document matches. By presenting a diverse range of negative examples, the model is capable of robust discriminative capabilities that improve its ability to distinguish high-quality matches from poor matches(Robinson et al., 2020).

In the context of contrastive learning, negative sampling plays a fundamental role in shaping the embedding space (T. Chen et al., 2020). When incorporated into triplet loss or InfoNCE loss functions, negative samples encourage the model to pull closer to the embeddings of relevant query-document pairs while pushing apart the embeddings of irrelevant pairs (Oord, Yazhe Li, and Vinyals, 2018). This objective creates more discriminative and rich embeddings, as the model learns to maximise similarity between positive pairs and minimise similarity between negative pairs (Khosla et al., 2020). The quality and diversity of negative samples significantly impact model performance, with hard negative examples that are semantically similar but ultimately irrelevant proving particularly valuable for learning fine-grained distinctions. For this study, a corpus of financial news articles, including stock price reports and organisational announcements, was selected as the negative sampling corpus. This choice is motivated by the contrastive linguistic and thematic distance between financial reporting and Holocaust testimony: the two domains share general English grammar but differ substantially in vocabulary, named entities, emotional register, and discourse structure, making them effective negative examples without introducing ambiguity.

Triplet Dataset Preparation

Preparing a triplet dataset allows the model to distinguish similar and dissimilar items effectively by simplifying data handling, enabling negative sampling, and converting data into embeddings (Schroff, Kalenichenko, and Philbin, 2015). As illustrated in figure 7.2 *TripletDataset* class, a custom dataset class included from PyTorch's Dataset library, was employed (Paszke et al., 2019) to work with triplet loss frameworks. Each data point is organised into three components: a query (anchor), a positive example, and a negative example, forming the basic structure for contrastive learning.

Triplet margin loss is a loss function widely used in metric learning and embedding learning tasks (Schroff, Kalenichenko, and Philbin, 2015). As illustrated in Figure 7.3, the primary objective is to learn embeddings such that similar examples are positioned closer together in the embedding space while dissimilar examples are pushed farther apart (Hoffer and Ailon, 2015). The loss function is defined as:

$$L = max(d(A, P) - d(A, N) + margin, 0)$$

where:

- Anchor (A): The reference input.
- Positive (P): A sample from the same class as the anchor.

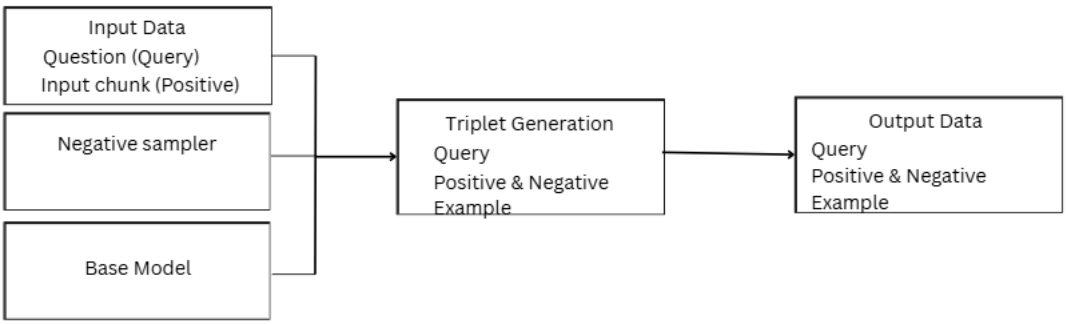

Figure 7.2: Synthetic Triplet Generation Approach for Adapter Training

- Negative (N): A sample from a different class.
- $d(.)$ is a distance metric (commonly Euclidean or cosine distance).
- $Margin$ is a hyper-parameter that enforces a minimum distance between the positive and negative pairs

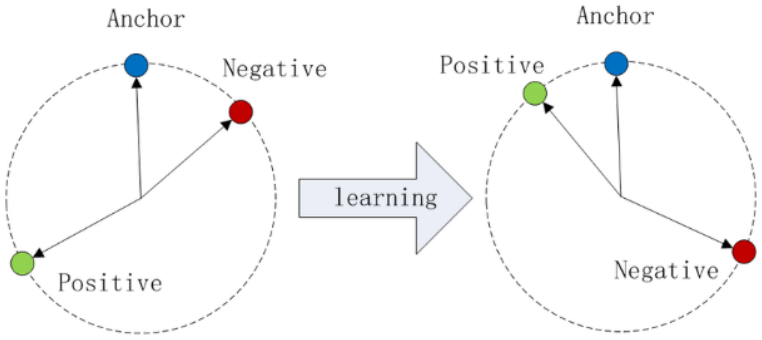

Figure 7.3: Triplet Loss Training Process

The loss encourages the model to learn embeddings where the distance between the anchor and positive should be smaller than the distance between the anchor and negative, by at least the margin.

$$d(A, P) < d(A, N) - margin$$

As a hyperparameter, triplet loss prevents embeddings from clustering too close, even if they are of different classes. It stabilises the training by avoiding trivial solutions where distances are all equal.

7.1.1 Corpus Preparation and Domain Adaptation

Since the proposed methodology requires query-document pairs that represent domain-specific information, we selected the 1783 oral narratives from our original, which is discussed in 3. Given the absence of pre-existing linguistic resources, we have employed an LLM to augment the queries and evaluated the output quality through human assessment. The question generation process is constrained to have WH-questions (such as "who", "when", and "what" queries), which are designed

to extract specific named entities and factual information from the testimony. The overall testimony was divided into chunks, and each question-answer pair was generated based on the information on that particular chunk. The primary goal for implementing these constraints was to minimise hallucination and generate information that is not explicitly included in the testimony. The suggested synthetic question-answer generation approach ensures that the training data is bound to the specific testimony. It mitigates the risk of the model relying on spurious correlations or hallucinations during retrieval. Figure 7.4 represents the prompt template used for the augmenting queries.

Synthetic queries were generated using the LLM model, where each testimony was chunked into blocks of 800 tokens with an overlap of 400 using the **sentence-transformers/all-MiniLM-L6-v2** model. Each testimony chunk is passed to an LLM (deepseek-chat) to generate contextually appropriate questions ($Queries = 20$) that the passage would answer. This approach addresses the problem of limited labelled training data in specialised domains while maintaining domain specificity. Further, it is able to test the same user query and assess it based on the retrieval/rank of the expected chunk. Across the corpus of 1,783 testimonies, they generated a total of 74,343 text chunks. These chunks were subsequently divided into training and validation sets using an 80/20 ratio.

7.1.2 Model Architecture and Training Pipeline

In this section, the implemented model architecture and training process are further explained. The proposed approach was developed using the Python language, the Sentence Transformers library for embedding generation, PyTorch for adapter training, and a standard vector database for similarity search. All the experiments were conducted using an A100 GPU environment. The code and training procedures were designed to be reproducible and adaptable to other specialised domains within digital humanities. The proposed method consists of three primary components:

Base Embedding Model: A pre-trained sentence transformer generates semantic

```
You are an AI assistant tasked with generating realistic, semantically rich
    ↪ questions that reflect user queries for information about named entities
    ↪ found in Holocaust oral testimonies.

Given: {chunk}

Instructions:
1. Read the document and identify named entities such as:
   - DATE, PERSON, ORGANISATION, COUNTRY, CITY, STREET, RIVER, FOREST, MOUNTAIN
   - CONCENTRATION CAMP, GHETTO, MILITARY RANK, SHIP, HAPPENING, DISEASE
   - ETHNICITY, NATIONALITY, LANGUAGE
2. Generate exactly 20 diverse user-style
    questions that could retrieve any of the entities found. Use natural phrasing
    ↪  including slang, typos, contractions, and both formal/informal styles.
3. Do NOT group questions by entity type. Just generate the full list.
4. Include the original input document ("chunk") in the output.
5. Ensure the question is semantically related to the named entities in the
    ↪ document content WITHOUT directly copying phrases.
6. Output must be valid JSON only, with no extra text or markdown.  Ensure proper
    brackets, quotes, and commas.

Return a JSON object with the following structure:
{
  "question_1": "Generated question text",
  "question_2": "Generated question text",
  ...
}
```

Figure 7.4: Prompt Template for Query Generation from Testimony Passages

embeddings, transforming both queries and documents into dense vector representations within a shared embedding space. The base model weights remain frozen throughout all training and inference stages.

Linear Adapter Layer: A trainable linear transformation matrix modifies query embeddings to better align them with the distributional characteristics of the domain-specific document corpus. The adapter acts as a learnt mapping that bridges the representational gap between general-purpose query formulations and specialised testimony content, without modifying either the base model or the pre-computed document embeddings. The query embedding Q is a dense vector of 384 dimensions ($Q \in \mathbb{R}^{384}$), produced by encoding the raw query string through the frozen base embedding model. Each dimension captures a component of the query's semantic meaning within the model's learned representation space, forming a single point in a high-dimensional semantic embedding space.

Although expressed compactly as a linear transformation, the adapter $\mathcal{A}$ is implemented as a bottleneck architecture ($384 \rightarrow 128 \rightarrow 384$) comprising a down-projection layer, a non-linear activation function, and an up-projection layer, together with a residual connection and L_2-normalisation to unit length. The parameters W and b in the notation $Q' = WQ + b$ collectively represent all learnable weights across these layers, and the expression serves as a schematic summary of the full transformation.

The bias term b is not set manually; it is learned entirely through training via backpropagation. Adapter parameters are optimised using the AdamW optimiser (Loshchilov and Hutter, 2017) with a learning rate of 10^{-4}, a linear warmup schedule of 100 steps followed by linear decay, and gradient clipping at a maximum norm of 1.0, over 10 epochs with a batch size of 32. The loss function combines InfoNCE Loss, which pulls adapted query embeddings closer to their ground-truth document chunk while pushing them away from in-batch negative examples using a temperature parameter of $T = 0.05$, with Triplet Margin Loss, which enforces a minimum margin of $m = 0.5$ between positive and hard-negative pairs weighted at $0.3\times$

as a supplementary contrastive signal. The resulting bias b encodes a domain-specific shift that systematically realigns query embeddings toward the subspace occupied by relevant testimony documents, thereby improving retrieval precision without modifying either the base embedding model or the pre-computed document embeddings.

Retrieval Mechanism: A cosine similarity-based retrieval approach compares adapted query embeddings with the pre-computed document embeddings to identify the most relevant passages.

7.1.2.1 Embedding Generation

Document embeddings were generated as a one-time preprocessing step. The testimony corpus was segmented using semantic chunking to preserve the contextual meaning across oral testimony transcripts. Each chunk was encoded using the embedding model, producing dense vector representations stored in a vector database (Chroma DB) for efficient retrieval. While computationally intensive, this preprocessing step is performed only once and does not require repetition during adapter training or subsequent model updates. Figure 7.5 illustrates the transformation of the query embedding space used in DsQoLA.

Four embedding models were selected to evaluate the linear adapter across architectures with varying size and task-specific optimisation:

- *all-MiniLM-L6-v2*: Small size and fast inference speed (base model).
- *all-MiniLM-L12-v2*: a more expressive 12-layer variant of the above.
- *multi-qa-mpnet-base-dot-v1*: Fine-tuned specifically for question-answering retrieval tasks.
- *stsb-roberta-base*: Model trained specifically on Semantic Textual Similarity data.

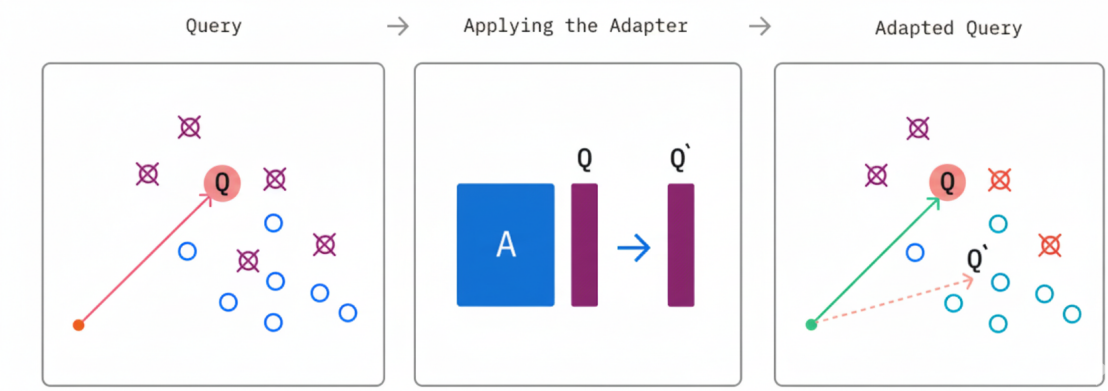

Figure 7.5: The original query embedding Q(left) is the base semantic space, whose nearest neighbours may not correspond to the relevant documents. The Linear Adapter A, parameterised by W and b, applies a transformation $Q' = WQ+b$ (middle), effectively realigning the query representation toward the subspace occupied by relevant documents. The adapted query Q' (right) shows higher similarity with relevant document embeddings, thereby improving retrieval precision without re-embedding the corpus.

7.1.2.2 Linear Adapter Training

The linear adapter is implemented as a bottleneck architecture with residual connections, designed to learn a domain-specific modification while maintaining computational efficiency via the following steps:

Forward Pass: Query strings from the training set were encoded using the frozen base embedding model, producing initial 384-dimensional query embeddings. These embeddings were then transformed through the linear adapter network comprising down-projection, non-linear activation, and up-projection layers with residual connection to generate adapted query representations that were subsequently L2-normalised to unit vectors.

Loss Calculation: The approach employed a hybrid loss function combining two complementary objectives: (1) InfoNCE Loss (Information Noise-Contrastive

Estimation), which treats all other samples within each training batch as negative examples, enabling the adapted query embeddings to calculate high cosine similarity with their corresponding positive document embeddings (ground-truth testimony chunks) while maintaining low similarity with in-batch negative examples, computed with a temperature parameter $T = 0.05$; (2) Triplet Margin Loss with hard negatives sampled from an external document corpus, enforcing a minimum margin ($m = 0.5$) between positive and negative similarities, weighted at $0.3\times$ to provide a supplementary contrastive signal without dominating the training objective.

Optimisation: Adapter parameters were optimised using AdamW (Loshchilov and Hutter, 2017) with a learning rate of 10^{-4}. A linear warmup schedule (100 warmup steps) followed by linear decay stabilised early training and improved convergence. Gradient clipping (max norm = 1.0) prevented exploding gradients. Training continued for 10 epochs with a batch size of 32. Throughout all stages, the base embedding model remained frozen, and document embeddings remained unchanged, ensuring that the document index never required recomputation after adapter training.

7.1.2.3 Retrieval Pipeline Integration

The trained linear adapter was integrated into an RAG pipeline to enable end-to-end query processing and response generation. The pipeline consists of four stages that transform user queries into contextually meaningful responses.

1. Query Processing: User queries are first encoded using the base embedding model to produce vector representations. As the next step, query embeddings are transformed through the trained linear adapter to better align with domain-specific information. This transformation allows to learn the domain knowledge during adapter training to enhance the semantic matching between queries and documents.

2. Similarity Search: The similarity search analyses the adapted query embeddings against pre-computed document embeddings stored in the vector

Algorithm 2 Linear Adapter Training

Notation: A_θ: adapter parameters, M_{base}: base model, P: passage embeddings, E: embeddings, $\mathcal{L}$: loss

Input: Training data D_{train}, base model M_{base}, negatives D_{neg}

Output: Trained adapter A_θ

Initialize adapter A_θ ($384 \to 128 \to 384$) and optimiser.

Pre-compute document embeddings: $P = \{M_{\text{base}}(c_i)\}_{i=1}^{N}$

for each epoch **do**

  for each batch **do**

    Sample and encode queries through M_{base}

    Encode hard negatives from D_{neg}

    Adapt queries: $E_{\text{adapted}} \leftarrow A_\theta(E_{\text{query}})$

    Compute InfoNCE loss (in-batch negatives)

    Compute Triplet loss (hard negatives)

    Combined loss: $\mathcal{L} \leftarrow \mathcal{L}_{\text{InfoNCE}} + 0.3 \times \mathcal{L}_{\text{triplet}}$

    Update θ via backpropagation with gradient clipping

  end for

end for

return A_θ

database. Cosine similarity serves as the distance metric for measuring semantic relatedness between query and document vectors. The system retrieves the top-k most similar document chunks, where k is a configurable parameter that balances retrieval breadth against computational efficiency and context window constraints.

3. Context Augmentation: In context augmentation, the retrieved document chunks are combined into a unified context string, which serves as the factual basis for response generation, integrating the retrieval and generation components of the pipeline. The concatenation process preserves the relevance ranking from the similarity search while packaging multiple information sources into a format suitable for language model consumption.

4. Response Generation: In the response generation, LLM processes both

the original user query and the augmented context to formulate responses. The LLM synthesises information from the retrieved chunks, ensuring that generated responses are grounded in the specific document rather than relying solely on parametric knowledge of the LLM.

7.1.3 Evaluation and Comparative Analysis

Evaluation focused on how accurately the adapter-enhanced retrieval system identifies the correct testimony chunk from the large corpus. Two standard information retrieval metrics were selected: Mean Reciprocal Rank (MRR) and Recall@K (Hit Rate).

- **Mean Reciprocal Rank (MRR)** (Y. Shi et al., 2012; Jadon and Patil, 2024): MRR measures the average reciprocal rank of the first correct result across all queries. It is particularly appropriate here because each query has a single ground-truth chunk:

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i}$$

  where

  - $|Q|$ is the number of queries
  - $rank_i$ is the rank of the first correct answer for the $i-th$ query

- **Recall@K** (Jadon and Patil, 2024): Recall@K, also known as Hit Rate, is the proportion of queries for which the correct chunk appears within the top-K retrieved results. With a single ground truth per query, this is a binary measure:

$$\text{Recall@k} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\mathbb{1}(rank_i \leq k)$$

  where:

  - $|Q|$ is the number of queries
  - $rank_i$ is the rank of the first correct answer for the $i-th$ query
  - $\mathbb{1}$ is the indicator function, which equals 1 if the condition inside the parentheses is true, and 0 otherwise

- k is the cut-off rank

The linear adapter was trained for different numbers of epochs (10, 30, and 40) using the hybrid loss function. Figure 7.6 illustrates the progression of training loss throughout the training process. The absence of increasing loss in the last epochs confirms that the dropout regularisation and bottleneck architecture effectively prevented overfitting.

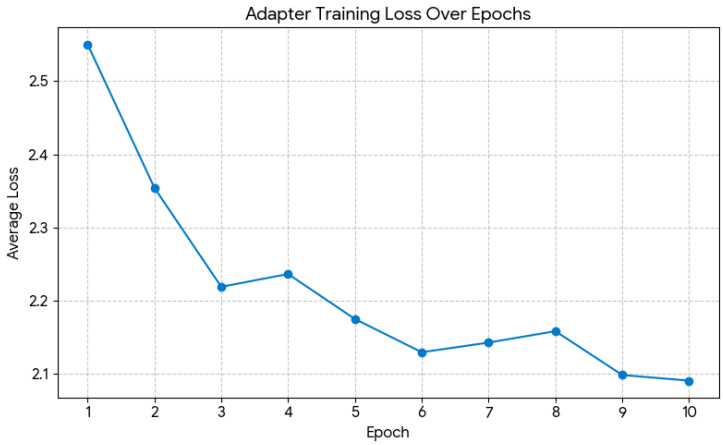

Figure 7.6: Average Training Loss over Epochs for DsQoLA

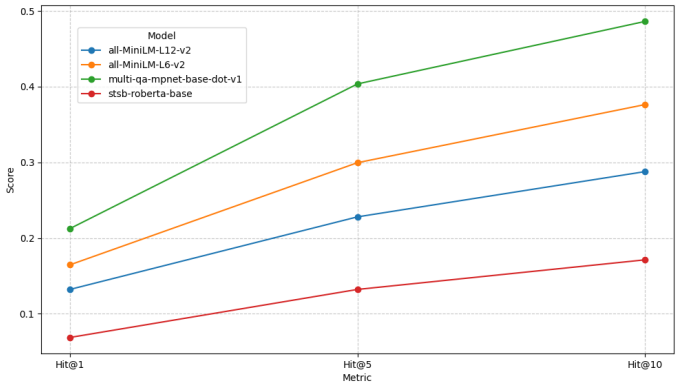

Figure 7.7: Overall Retrieval Performance Evaluation with and with the Linear Adapter

Figure 7.6 presents the adapter's training loss over 10 epochs, demonstrating

the optimisation process. The loss decreased from 2.54 to 2.09 (17.7% reduction), with a prompt initial descent in epochs 1-3 followed by gradual stabilisation. The smooth convergence pattern without significant oscillations indicates stable learning dynamics and begins distinguishing between relevant and irrelevant testimony chunks. The absence of increasing loss in the last epochs confirms that the dropout regularisation and bottleneck architecture prevented overfitting effectively.

As shown in Table 7.1 and Figure 7.7, the linear adapter consistently improved retrieval performance across all four base models and all metrics. The adapter-enhanced **multi-qa-mpnet-base-dot-v1** achieved the best absolute scores across all Hit@K and MRR metrics, reflecting its task-specific pre-training for question-answering retrieval. Conversely, the adapter applied to **stsb-roberta-base** produced the largest relative gains (over 63% improvement) in Hit@1 and nearly (60% improvement)in MRR, suggesting that the adapter is particularly effective at compensating for weaker base models whose general-purpose embeddings are least suited to domain-specific retrieval. This finding is practically significant because it demonstrates that DsQoLA can meaningfully enhance retrieval quality even when only lightweight or resource-constrained base models were used. Moreover, having a lightweight approach is essential in digital humanities research contexts where access to high-performance computing is not guaranteed.

7.2 Challenges of Extraction

Although promising results were achieved through the Domain-specific Query-only Linear Adapter (DsQoLA) approach, several significant challenges were observed during the data preparation, model training, evaluation, and deployment stages.

The absence of labelled query-document pairs in the Holocaust domain necessitated the generation of synthetic training data. However, extensive validation is necessary to ensure that the LLM-generated queries remain semantically grounded in the actual testimonies, without introducing hallucinated information. Furthermore,

Table 7.1: Performance comparison of models on retrieval metrics.

Model	**Metric**	**Base-Score**	**Adapter-Score**	Δ †	Δ %‡
all-MiniLM-L6-v2	Hit@1	0.1395	0.1646	0.0251	18.0
	Hit@5	0.2280	0.2996	0.0716	31.4
	Hit@10	0.2863	0.3764	0.0901	31.4
	MRR	0.1780	0.2233	0.0453	25.4
all-MiniLM-L12-v2	Hit@1	0.0974	0.1321	0.0347	35.6
	Hit@5	0.1749	0.2280	0.0531	30.4
	Hit@10	0.2162	0.2878	0.0716	33.1
	MRR	0.1310	0.1731	0.0421	32.1
multi-qa-mpnet-base	Hit@1	0.1860	**0.2125**	0.0265	14.3
	Hit@5	0.3240	**0.4037**	0.0797	24.6
	Hit@10	0.3904	**0.4863**	0.0959	24.6
	MRR	0.2435	**0.2903**	0.0468	19.2
stsb-roberta-base	Hit@1	0.0421	0.0686	0.0265	63.2
	Hit@5	0.0856	0.1321	0.0465	54.3
	Hit@10	0.1122	0.1712	0.0590	52.6
	MRR	0.0595	0.0949	0.0354	59.6

† Absolute Improvement (Δ) = Adapter score − Base score.

‡ Relative Improvement (Δ %) = $(\frac{\text{Adapter Score}}{\text{Base Score}} - 1) \times 100$.

Table 7.2: Consolidated Multi-Model Comparison: Absolute and Relative Improvement

the additional constraint of incorporating WH-questions into the prompt design helped mitigate this risk. However, human assessment was required to verify output quality, creating a labour-intensive validation bottleneck.

The linear adapter was explicitly trained on Holocaust testimony data with synthetically generated queries following particular linguistic patterns. However, questions remain about how well this approach generalises to other specialised digital humanities corpora with different characteristics, such as cultural contexts, literary texts with varying narrative structures, or historical documents with distinct temporal and linguistic features. Each new domain may require retraining with domain-specific synthetic data, limiting the transferability of the approach. However, eventually that challenge became the spotlight of this approach, which involves training a lightweight adapter for individual domains rather than developing a generalised, larger model.

Although the linear adapter approach is significantly more efficient than full model fine-tuning, the training process still requires substantial computational resources. For example, training multiple epochs with large batch sizes with the hybrid loss function requires A100 GPU resources, which are not available to all researchers working in digital humanities. Additionally, the need to experiment with multiple base embedding models to identify optimal performance required repeated training cycles, multiplying resource demands. Furthermore, incorporating the trained linear adapter into the RAG pipeline necessitates careful coordination of multiple components, including the frozen base model, the adapter transformation layer, the vector database, and the generative LLM. This tightly coupled and modularised architecture intensifies challenges in model reproducibility and system maintenance, where updates to any individual component require a comprehensive re-evaluation of the pipeline. These challenges emphasise that, while the DsQoLA approach is able to improve the performance in domain-specific retrieval, successful implementation necessitates careful consideration of data quality, computational resources, evaluation methodologies, and deployment contexts. Future work should

address these limitations to enhance the robustness and applicability of the approach across diverse digital humanities applications.

7.3 Chapter Summary

This chapter presented a comprehensive approach to enhance retrieval accuracy in domain-specific Retrieval-Augmented Generation (RAG). The research addressed the fundamental challenge that general-purpose embedding models, trained primarily on web-based text, often fail to capture the domain-specific information. Therefore, this chapter detailed the complete retrieval pipeline integration, explaining the procedure of trained adapter transforms user queries through four sequential stages: query processing, similarity search, context augmentation, and response generation. The DsQoLA contributes to the broader field of information retrieval in digital humanities by demonstrating that domain adaptation need not require extensive computational resources or large-scale model retraining. Instead, using parameter-efficient approaches could achieve substantial improvements in retrieval accuracy, enabling more effective access to specialised historical and cultural corpora. Furthermore, the success of this approach with Holocaust testimonies suggests its potential applicability to other low-resource domains, including literary archives, ethnographic recordings, and other oral history collections.

In conclusion, this chapter established that the combination of synthetic query generation, combined with lightweight adapter architectures, provides a viable pathway for enhancing retrieval performance in domain-specific RAG pipelines. The proposed approach balances computational efficiency with performance gains, offering a practical solution for researchers who are working towards improving information access in specialised digital humanities collections. To address identified challenges, future work will focus on improving generalisation, enhancing multilingual support, and developing context-sensitive evaluation metrics.

Chapter 8

Ethics of LLMs in Processing Oral Historical Narratives

Whoever listens to a witness becomes a witness.

Elie Wiesel(Survivor)

8.1 Introduction

In recent decades, technology has simplified complex tasks across multiple disciplines, from the sciences to the digital humanities. As discussed in the above chapters, traditional machine learning algorithms have been widely employed for historical document analysis. The emergence of generative AI has introduced new ethical dilemmas when applied to the DH domain. LLMs are not neutral tools; they function on statistical principles which are fundamentally different from humanistic interpretation. As a result, over time, a set of epistemological and ethical challenges that relate to the custodians of human narrative are raised by forcing a re-evaluation of core principles such as consent, authenticity, authority, and the preservation of context.

The early days of digitisation focused on preparing recordings from physical tape to digital files, creating passive, more accessible repositories of oral histories. The

current integration of LLMs creates a different direction to explore, from archive to algorithm. This transition goes beyond digital storage; it subjects personal narratives to computational analysis, where they are redefined as data for training algorithms, tokenised sequences, and sources for probabilistic inference. Such a transformation necessitates an ethical framework for preserving narrative content and identifying challenges of its reinterpretation.

Therefore, this chapter focuses on the identification of ethical concerns involved in processing sensitive historical documents using generative AI, mitigating strategies for the technical and ethical risks associated with LLMs and potential benefits and future research directions of employing generative AI in the context of sensitive data. Additionally, this chapter examines the risks of misrepresentation, exploitation, and harm that arise when the human voice is processed as data. Considering Holocaust survivor testimony as a central case study, it explores the broader ethical and procedural implications of applying generative AI to historically significant and trauma-induced content.

8.2 Traditional Ethical Pillars in Oral History

The oral history interviews are a dynamic, interpersonal exchange where the interviewer and narrator collaboratively construct the narrative. The interviewer's questions, presence, and relationship directly shape the story that emerges (Morrissey, 2002). Consequently, the interview yields a subjective truth of lived experience and memory, not an objective, verifiable record. This foundational characteristic presents a primary challenge for LLMs, which process the resulting transcript in isolation. Beyond their content, personal and community narratives often function as the representation of identity. They are constitutive of cultural history, values, and collective memory and are frequently considered proprietary (Kekki, 2024). Furthermore, the textual content of an oral narrative is only one component of its meaning. Equally critical are the paralinguistic features such as tonal

inflection, rhythmic pacing, emotional resonance (like a voice breaking with sorrow or lifting with laughter), and deliberate silences. These elements carry the affective and emphatic weight of the testimony, interpreting nuance, sarcasm, sorrow, and emphasis, where general LLMs are inherently blind to analysing the above features (Bojić et al., 2025).

Oral history is built upon a robust ethical framework, shaped by decades of practice and a profound understanding that oral narratives are not merely data points but deeply personal expressions of lived experience. As a result, the field has developed ethical pillars centred on respect, reciprocity, and responsibility (Shopes, 2007). These principles, which include informed consent, stewardship (vs ownership), and contextual fidelity, safeguard the dignity and rights of narrators and uphold the authenticity and trustworthiness of the historical record itself. As this research examines the integration of LLMs into oral history, these established ethical standards provide the essential benchmark against which the promises and perils of AI-driven approaches must be critically assessed. This analysis reveals that the fundamental operational logic of LLMs is directly tied to the core ethical commitments of oral history.

In oral history, informed consent ensures the narrator understands and agrees to how their story will be used (Shopes, 2007). This creates a fundamental problem with AI: a standard consent form cannot meaningfully cover having a story processed by an incomprehensible "black box" used to train other systems or repurposed to generate new content. Since traditional consent originally obtained at the time of recording (pre-LLM era) cannot anticipate these unforeseen algorithmic uses, most existing practices are ethically unsuitable for LLM projects. The ethical principle governing oral histories is one of stewardship; archivists and researchers are temporary guardians of a narrative, bound by a relationship-based ethic of care. LLM development, in contrast, often operates on a logic of data ownership and extraction. Applying this logic severs the sacred bond of trust, transforming the researcher into a "data extractor" and the narrator's story into a commodity for a

distant third party. Oral historians maintain that a narrative is meaningless without its historical, interview, and biographical context, and they are ethically bound to preserve its nuances (Charlton, Myers, and Sharpless, 2007; Beard, 2017). LLMs, however, are inherently decontextualising engines, where the text is extracted from its original context and often "hallucinates" or summarises in ways that distort the original meaning, introduce factual errors, or oversimplify complex truths (Anh-Hoang, Tran, and L.-M. Nguyen, 2025). The integration of LLMs into oral history is not a straightforward methodological enhancement but a development that poses a fundamental ethical challenge. The core operational principles of LLMs, including their 'black box' nature, data processing, and propensity to decontextualise, directly conflict with the foundational ethical pillars of oral history: informed consent, relational stewardship, and contextual integrity

8.3 Challenges Inherited from Generative AI

Unlike traditional methods of AI, which are often limited to analysing or classifying existing data, generative AI can create new content. These models are trained on diverse internet data, enabling them to produce content that includes natural language text, images, videos, music, and software code. While techniques such as prompt engineering and Retrieval-Augmented Generation (RAG) introduce degrees of human oversight, shaping inputs or grounding outputs in curated external sources, the generative process itself remains fundamentally autonomous: LLMs produce responses by statistically predicting token sequences, with no guarantee of factual accuracy or contextual fidelity at the point of generation. This process carries inherent risks when generating content related to historically sensitive and factually accurate topics. Consequently, in UNESCO's 2024 report (Makhortykh, 2024), several concerns were raised regarding the influence of generative AI on the Holocaust domain:

- Manipulating and leveraging AI models to produce and spread hate speech

- False statements and narratives in generative AI content
- Producing fake historical evidence
- Jeopardising belief in authentic historical evidence
- Oversimplifying Holocaust histories
- Language bias: reinforcing gaps in global Holocaust understanding

These concerns are not merely hypothetical. The Anti-Defamation League (Anti-Defamation League, 2025) conducted the most comprehensive empirical evaluation to date of antisemitic bias in leading LLMs, querying GPT (OpenAI), Claude (Anthropic), Gemini (Google), and Llama (Meta) across 34,400 responses. The study found that all four models exhibited measurable anti-Jewish and anti-Israel bias and demonstrated a systematic inability to reject antisemitic tropes and Holocaust conspiracy theories (Anti-Defamation League, 2025). Llama was identified as exhibiting the most profound bias among open-source models. These findings directly substantiate the UNESCO concerns above with observable, quantified test results from the specific models most commonly deployed in digital humanities research.

Although generative AI systems are not designed to promote hate speech or offensive content, they may nevertheless produce biased or harmful material due to the nature of their training data. Recent model iterations attempt to detect and block offensive content; however, misinterpretations persist (Afreen, Mohaghegh, and Doborjeh, 2025). Users may misconstrue these outputs as intentional bias, when in fact they arise from the model's statistical learning processes rather than deliberate human input. For instance, the generation of Holocaust denial statements illustrates how a model's exposure to unreliable or extremist material during training can yield historically inaccurate and deeply harmful outputs (Dutta et al., 2024). Because developers often scrape data indiscriminately from the internet without expert curation, these issues are neither rare nor restricted to Holocaust discourse, they pervade other sensitive social, historical and cultural domains as well (Timmons et al., 2023; Tiribelli et al., 2024).

LLMs tend to oversimplify history, which poses a specific risk when addressing events like the Holocaust. Because LLMs are trained on generalised data, they may inadvertently dilute the event's specific context and horror. Therefore, it is crucial to treat AI-generated content on this topic with caution and always cross-reference it with authoritative historical archives when extracting and interpreting sensitive and critical information.

Statistical Prediction vs. Historical Testimony: LLMs predict word sequences based on probabilistic patterns identified from training data, and their performance is often measured in terms of linguistic coherence rather than factual accuracy. However, oral history derives its authority from lived experience: it is a phenomenological act of bearing witness, where truth is grounded in memory, rather than statistical likelihood. Because of this fundamental difference, two significant risks arise as follows:

- Marginalisation of the implausible: Oral history's most significant contributions often lie in unique, counter-hegemonic accounts. However, the model's architectural bias toward common patterns systematically marginalises these narratives, silencing the very testimonies that challenge established historical paradigms (Agarwal, Naaman, and Vashistha, 2025).

- Conflation and fabrication of history: In pursuit of textual coherence, an LLM may "hallucinate" plausible-sounding but entirely fictional details of the Holocaust. While statistical consistency signifies success for the model, it constitutes a profound distortion of the historical record for the historian, for whom accuracy and fidelity are paramount.

Decontextualisation Through Tokenisation The tokenisation step in LLMs, where human language is transformed into numerical tokens, can cause decontextualisation. This process detaches narratives from the contextual setting that gives them meaning such as the interpersonal setting of the interview, the relationship between interviewer and narrator, cultural codes, and the socio-political conditions of the

recording moment. The consequences of this decontextualisation are significant and can be observed in several ways:

- Loss of situated meaning: Sarcasm, irony, trauma, and culturally embedded references can be wholly misinterpreted when processed in isolation from their original context.

- The illusion of objectivity: Representing testimony as a token sequence represents a false sense of neutrality, hiding the fact that it is always co-created and interpreted. This simplified view treats oral testimony as mere data to be extracted, rather than as something that carries ethical responsibility and requires careful stewardship.

Standardisation of Unique Human Voices LLMs are good at finding patterns in language and producing consistent results. However, this strength becomes a problem when dealing with oral testimony's unique and personal nature. The individual qualities of a narrator's voice—such as dialect, sentence structure, rhythm, pauses, and emotion—are not background noise to be removed but essential parts of meaning. They express personality, feeling, and cultural identity. When an algorithm tries to make language uniform, it can cause harm by erasing these differences and reducing the richness of human expression. This standardisation manifests in several key ways:

- Erasure of idiosyncrasy: When summarising or editing transcripts, LLMs may correct informal speech or remove repeated phrases, which can take away the individuality of a person's voice.

- Homogenisation of experience: When combining many different accounts, the model often merges them into one simplified version, losing the diversity of voices that oral history aims to preserve. Further, models favour a single, general story instead of recognising the complex and varied realities of people's lived experiences.

These issues are not just technical problems but deep philosophical challenges. They show that using LLMs in oral history is not neutral or purely analytical; it actively changes the testimony by applying the processes of prediction, decontextualisation, and standardisation. Without sustained critical oversight and ethically informed mitigation strategies, the use of generative AI in this domain risks eroding the narrative integrity, contextual fidelity, and singular human voice that form the ethical and epistemic foundations of the oral history tradition.

8.4 Ethical Challenges Arising from Technical Limitations of LLMs

The use of LLMs to process oral histories introduces ethical challenges that traditional research ethics frameworks are not fully equipped to handle. Historically, ethical discussions have centred on how humans collect, interpret, and share personal narratives. However, LLMs operate in fundamentally different ways of using highly complex, continuously updated data, and rely on data-driven processes that are often opaque and beyond direct human oversight. These technical characteristics not only reshape how knowledge is stored and disseminated but also challenge established understandings of consent, ownership, and privacy. This section explores how the technical limitations of LLMs intersect with core ethical principles, focusing on informed consent, cultural sovereignty, historical accuracy, and personal dignity.

8.4.1 Consent and ownership

Standard oral history consent forms are structurally designed for archival use and potential research by human scholars, assuming that any subsequent research will be bounded by established timeframes, contexts, and explicit research purposes. This model is inadequate for the processing of narratives by LLMs, a mechanism that is inherently open-ended, non-transparent, and continuously subject to algorithmic

change. Consent given for a human to listen or read does not equate to consent for a machine to ingest, analyse, and repurpose that narrative indefinitely. This renders initial consent insufficient and potentially invalid. Although explicit and informed consent is required, the complexity of LLMs makes it impossible to ensure that narrators truly understand their implications (how their data becomes weights in a model, what "training" entails, or the potential for generating synthetic output). This imbalance of knowledge invalidates genuine consent, reducing it to a procedural checkbox instead of a meaningful ethical commitment. Ethical research respects the participant's right to withdraw their data at any stage. However, this right becomes functionally obsolete when narratives are used to train an LLM, because the information in the narrative will be statistically absorbed into the model's parameters, which cannot be "deleted" or "unlearned" like a file from a server without any chance of being recovered. This technical limitation further confirms that once a human's voice and speech patterns have been used to train an AI, it is impossible to completely and permanently remove that personal data. These challenges around consent and data removal reveal a deeper ethical concern: human narratives are increasingly being treated as machine-readable assets rather than personal or cultural expressions. This issue leads directly into questions of agency, exploitation, and cultural sovereignty, where power imbalances become even more pronounced.

8.4.2 Agency, Exploitation, and Cultural Sovereignty

In traditional oral history, the goal is to honour the human voice by allowing people to tell their own stories in their own way. However, when these stories are processed by large language models (LLMs), they are converted into numerical data and statistical patterns that computers can interpret. This transformation turns a person's lived experience, along with their emotions, culture, and personal meaning, into raw data. As a result, the storyteller, or narrator, loses some control over how their story is understood or used. Rather than being recognised as a person

sharing knowledge, they become a data source—an object to be analysed. This shift diminishes human agency and risks dehumanising the storyteller.

The modern form of extraction of data related to cultural knowledge, genetic information, or personal narratives (neo-colonial data extraction) highlights the global power imbalances inherent in how data is collected and used (J. S. Roberts and Montoya, 2022; Lynch et al., 2023). Many communities, particularly indigenous groups and marginalised people, share cultural knowledge or personal stories that are subsequently appropriated by powerful institutions in wealthier countries, such as major technology companies or elite universities. These institutions extract this data, process it using AI, and profit from the outcomes, whether through commercial products, academic research, or enhanced AI systems. Yet the original contributors, the communities whose knowledge formed the foundation of this work, typically receive neither benefit nor recognition. This dynamic mirrors the historical logic of colonialism, in which resources were taken from less powerful groups for the enrichment of dominant powers; today, the extracted resource is data rather than land or raw materials. Current intellectual property laws were designed for creations such as books, songs, and inventions which were not created using AI based on huge knowledge bases. This mismatch raises difficult questions: if an LLM is trained on sacred stories from the Holocaust and then generates a new version of those stories, who owns the resulting work? Is it the community that holds and stewards the original traditions or the developers who created the AI? When AI outputs draw from cultural heritage, yet the originating community has no control over or recognition for that use, it can amount to a new form of cultural appropriation, one in which technology effectively claims ownership over cultural expression. By reducing human experiences to data and stripping away cultural ownership, LLMs perpetuate structural injustice. However, the ethical crisis is not limited to these large-scale issues of exploitation. There are also profound personal and psychological dangers when an individual's private narrative is mishandled.

8.4.3 Accuracy, Misrepresentation, and Hallucination

Malfunctions and unintended behaviours in LLMs can lead to significant ethical breaches when processing sensitive and unstructured data. In the context of historically significant documents such as Holocaust survivor narratives, the use of LLMs requires heightened caution because they contain deeply personal and immutable information. It is imperative that LLM outputs maintain historical accuracy and ethical integrity without misrepresentation or distortion. Below are the concerns that tend to establish ethical breaches when using LLMs.

Bias in training data represents a significant ethical risk in generative AI, particularly when processing sensitive unstructured texts. Since most generative AI models are trained on large datasets scraped from the internet, they often reflect the biases, inaccuracies, and cultural insensitivities present in those sources, which may conflict with the real sources. Therefore, an accurate representation is required for historical archives. Misrepresentation of sensitive historical narratives and reproducing or amplifying antisemitic, racist, or discriminatory language could be considered possible ethical breaches related to the biases of the training data (Pearson, Seliya, and Dave, 2021).

Hallucination is a well-known issue in generative AI models, where the models produce information that is not based on actual data, leading to fabricated or inaccurate outputs. This is particularly concerning when generative AI is applied to sensitive and factual domains, as the AI may alter or misrepresent essential facts. Generative AI models, by their nature, often synthesise content based on patterns in their training data, which may lead to the unintentional creation of plausible-sounding but entirely fictional information. Generative AI models create human-like content, but their developers emphasise that the outputs are not always 100% accurate and reliable. Disclaimers typically warn users not to treat the generated results as the absolute truth without verification. (Venkit et al., 2024). Empirically, hallucination is not a rare edge case. Peer-reviewed studies have measured hallucinations in 31.4% of real-world LLM interactions, rising to 60% in

complex or specialised domains (Ren, Gruhlke, and Lauscher, 2025). Open-ended generation tasks demonstrate hallucination rates of 40–80%, the highest of any task category. Crucially, (Kalai and Vempala, 2024) demonstrated mathematically that eliminating hallucination in LLMs is not merely difficult but provably impossible given their generative architecture: any system that produces text by predicting probable token sequences will, by mathematical necessity, sometimes generate outputs not grounded in fact. For Holocaust testimonies, where a single fabricated detail can distort a survivor's account or introduce historically false information into the scholarly record, this is not a theoretical risk but a structural property of the technology. Ethical concerns arise through the information extracted through LLMs could spread misinformation, which could undermine the trust in the historical records stored in the archives and libraries.

LLMs are trained on general-purpose corpora designed to perform a wide range of NLP tasks. However, extensive training often leads to overgeneralisation, where the model applies generic patterns and assumptions to all contexts without recognising the distinct historical, cultural, and ethical significance of specific events, such as the Holocaust. As a result, the gravity of historically significant events such as the Holocaust may be diluted, neutralised, misrepresented, or oversimplified. The inability to capture event-specific nuances undermines scholarly interpretation and may cause ethical harm by diminishing the gravity of sensitive events, leading to inappropriate or insensitive outputs. The documents related to the Holocaust were recorded in various languages, reflecting the diverse linguistic backgrounds of survivors across Europe. These narratives often include native terms and region-specific expressions, which are crucial for preserving authenticity and cultural context. Although recent LLMs such as GPT-4, LLaMA 3, and BLOOM have substantially expanded their multilingual coverage, empirical studies consistently confirm that performance remains closely tied to a language's representation in the pre-training corpus: high-resource languages such as English, German, and French continue to benefit from significantly stronger model capabilities than lower-

resource European and non-European languages (Al Nazi, Hossain, and Al Mamun, 2025; W. X. Zhao et al., 2023). For Holocaust testimonies recorded in Yiddish, Hungarian, Polish, Czech, and other languages that are under-represented in modern training corpora, this imbalance translates directly into inaccurate translations and misinterpretations of culturally specific terminology.

8.4.4 Privacy, Dignity, and the Risk of Harm

The integration of LLMs in processing sensitive historical data, such as Holocaust testimonies, raises critical concerns around data privacy and security. Improper and insecure handling of such texts can lead to unauthorised access, data leaks, and breaches of confidentiality when the data includes personally identifiable information (PII) of survivors, victims, or their families. Moreover, without clear data ownership acknowledgement or institutional accountability, the risk of ethical violations increases significantly. These vulnerabilities compromise the ethical handling of survivor narratives but may also violate data protection regulations such as GDPR (Shopes, 2007). Given these testimonies' deeply personal and often traumatic nature, it is imperative to adopt robust encryption protocols, secure infrastructure, and ethical data governance frameworks to ensure that sensitive content is preserved and protected with the utmost care without offending any community in society. When people share their stories, researchers often try to protect their privacy by anonymising their names or other identifying details. However, this approach is not always effective with the emergence of AI. The statistical nature of LLMs means they do not 'forget' the data they are trained on; they learn latent patterns and correlations. As a result of that, there is a possibility of reconstructing identities or inferring sensitive attributes (like location, gender, or health status) from the subtleties of the narrative itself—turns of phrase, referenced events, or cultural context—rendering superficial anonymisation useless and violating the contextual integrity of the testimony.

LLMs can use powerful pattern recognition to infer an individual's identity, even

from seemingly minor clues such as a reference to a local event, a rare profession, or a distinctive personal experience (Smith et al., 2023). This risk has been quantified in practice. (Staab et al., 2024) demonstrated that GPT-4 can infer sensitive personal attributes such as location, occupation, and health status — from anonymised text with up to 85% accuracy, using only linguistic patterns and contextual cues. Applied to Holocaust testimonies, which contain highly distinctive personal details such as specific camp names, transport dates, and rare occupations, re-identification risk is substantially higher than in general text corpora. Even more troubling, LLMs could draw conclusions about information the narrator never explicitly stated. For instance, a model might deduce someone's health conditions, political views, or personal relationships based on subtle cues in their language. As a result, even anonymised data may not be truly private, because AI systems can reconstruct personal identities or sensitive details that were intended to remain hidden. Further, general-purpose LLMs often lack deep contextual reasoning, resulting in oversimplified, decontextualised, or factually inaccurate outputs, as they are typically trained on broad, general-purpose datasets that do not emphasise historical accuracy or trauma-informed language. They might not differentiate between factual historical content and fictional narratives or recognise the gravity of traumatic events. Such limitations can have an impact on historical significance.

Many oral histories include painful or deeply personal memories, such as stories about violence, loss, discrimination, or trauma. When such narratives are processed by an LLM, the machine treats them as just another piece of data, lacking emotional understanding or empathy. As a result, the AI might generate cold or detached summaries of someone's suffering, or worse, reproduce fragments of a traumatic story in a new context where others can see it. This can lead to re-traumatisation, causing the individual to experience emotional distress again because their story was handled carelessly or shared publicly. In human-centred research, an interviewer would approach these topics with compassion and ethical care, but AI cannot replicate that sensitivity. Together, these ethical issues show that while LLMs are powerful

tools, they also make existing moral and intellectual problems. Addressing these issues requires not only technical safeguards but also a rethinking of research ethics that prioritises dignity, context, and cultural responsibility.

8.5 Risk Mitigation Strategies for Responsible Use of LLMs

History, being a collection of narratives about the past, constructed by historians and aided by selected informants and interviewees, will always entail a degree of interpretability and subjectivity and is arguably not entirely objective in nature (Crane, 2006). However, as custodians of these narratives, historians and humanities scholars must take great care to ensure that histories are not unfaithfully recreated or distorted by any influence of technology. Therefore, when analysing historical texts using LLMs, careful consideration of ethical standards is required to avoid offending or misrepresenting any societal groups. As previously discussed, the technical limitations of LLMs can give rise to ethical risks. From this section, the following mitigation strategies were proposed to address potential ethical issues related to LLMs when handling historically specific, sensitive and unstructured data. However, generative AI models cannot reason like humans; they solely depend on the patterns recognised during the training process of language models.

- Fine-tuning domain-specific LLMs: Since most LLMs were trained on general-purpose data, they often lack the specificity and sensitivity required to generate accurate, contextually appropriate facts for the Holocaust domain. This can lead to ethical breaches, such as misrepresentation of historical events, oversimplification of survivor narratives, or perpetuation of biases, potentially undermining the dignity and authenticity of Holocaust testimonies. To mitigate these risks, domain-specific fine-tuning on Holocaust documents, including survivor testimonies, diaries, and official archives, is essential to enhance factual accuracy and cultural sensitivity. Additionally, to ensure

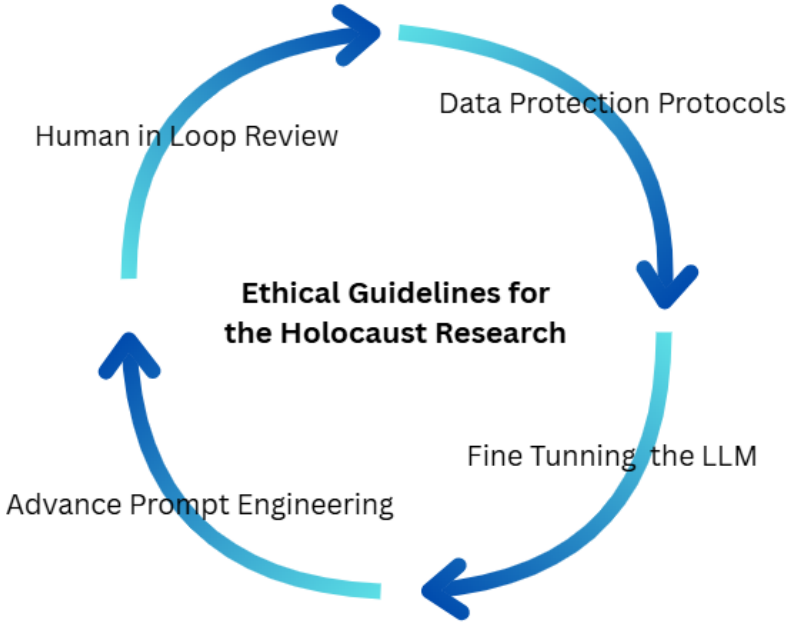

Figure 8.1: Integrated Mitigation Framework for Ethical LLM Deployment on Holocaust Testimonies

accurate handling of multilingual data relevant to the Holocaust, LLMs should incorporate multilingual training data in the process of fine-tuning to improve context-aware translation models and prevent misinterpretation. Moreover, the Retrieval-Augmented Generation (RAG) framework provides a concrete and technically robust architecture to integrate domain-specific information with a general-purpose LLM model instead of relying solely on a model's static, pre-trained knowledge, which may be incomplete, outdated, or lack the necessary nuance for sensitive domains. In the RAG approach, relevant information was dynamically retrieved from trusted, external knowledge bases before the LLM generated a response. The model is then prompted to ground its answer strictly in this retrieved evidence and tries to mitigate the ethical risks of hallucination and bias directly. Beyond RAG, other advanced techniques offer complementary pathways.

- Use context-enriched prompts combining with external knowledge sources: To

mitigate the ethical concerns associated with using LLMs in the Holocaust domain, such as misrepresentation, bias, and insensitivity, it is critical to employ context-enriched prompts integrated with external, domain-specific knowledge sources. These prompts are crafted to provide LLMs with detailed historical context, including specific instructions to prioritise factual accuracy and cultural sensitivity when generating content related to Holocaust survivor testimonies. As a part of the prompt, domain-specific external knowledge sources such as curated historical databases, glossaries of Holocaust terminology, and verified archival materials could be attached to enhance the model's ability to produce responses that reflect deep historical understanding. Teaching a model to extract deep contextual meanings from the prompts reduces the risks of misinterpretation, propagation of biased and inappropriate content, and oversimplification of traumatic narratives.

- Human-in-the-Loop Review: Human-in-the-Loop (HITL) review is a critical mechanism for mitigating ethical concerns when deploying LLMs related to historically sensitive domains such as Holocaust research. Given that LLMs, trained on general-purpose datasets, which may produce biased, inaccurate results, it is recommended to incorporate human oversight by domain experts, such as historians and linguists, to ensure that generated content aligns with the ethical, factual, and emotional dimensions of the Holocaust. HITL review involves expert validation on generated content for factual accuracy, cultural appropriateness, and ethical acceptability, to filter out distortions, anachronisms, or potentially offensive interpretations. Including iterative feedback loops is more effective, where experts refine model behaviour by adjusting prompts, reviewing outputs, and contributing to targeted fine-tuning using carefully curated Holocaust-specific corpora. Additionally, aligning model responses with verified historical archives strengthens the credibility and integrity of the content produced. By embedding HITL mechanisms into the model development and deployment pipeline, researchers could uphold ethical

standards, protect the authenticity of survivor testimonies, and prevent harm or misrepresentation.

- Data protection protocols - When using LLMs for sensitive domains, robust data protection protocols are critical to safeguarding the privacy and integrity of documents, such as survivor testimonies, personal diaries, and official records. Inadequate data security measures can lead to ethical breaches, including unauthorised access, data leaks, or misuse of personal information in the narratives, which could violate survivor dignity and erode trust in historical research. Additionally, most testimonies belong to institutions and archives, necessitating proper authorisation and ethical approval before their use to respect intellectual property rights and institutional agreements. To address these risks, comprehensive protocols are required to establish as end-to-end encryption for data storage and processing, restricting access controls and anonymisation techniques to remove personally identifiable information before analysis. Compliance with international data protection regulations, such as GDPR, is essential, particularly for multilingual testimonies spanning multiple jurisdictions. Additionally, integrating audit trails to track data usage and employing secure cloud environments can further enhance protection.

Based on our observations, the ethical challenges associated with deploying LLMs in historically and culturally sensitive domains necessitate a multifaceted, integrated approach. Our experiments have successfully adapted and validated four techniques described above and developed a robust framework for mitigating the risks of hallucination, bias, misrepresentation, and privacy violations in Holocaust-related applications.

The foundation of our approach rests on domain-specific fine-tuning and is augmented with retrieval-augmented generation (RAG). By fine-tuning LLMs on carefully curated Holocaust survivor testimonies, the model has been able to improve factual accuracy and cultural sensitivity. The RAG framework further improves the dynamic retrieval of verified information from trusted knowledge bases and

ensures the generated responses are grounded in authoritative sources rather than relying on pre-trained knowledge. Adapting these approaches enabled the risk of hallucination to be addressed while maintaining the nuance necessary for sensitive historical content.

Prompting is a key element in the LLM ecosystem. The use of context-enriched prompts, integrated with domain-specific external knowledge sources, has proven essential for guiding the model to produce reliable responses. By providing detailed historical context and explicit instructions with factual accuracy and cultural sensitivity and combining them with domain-specific dictionaries and trusted archival materials, LLMs could generate responses without biased interpretations and oversimplified survivor narratives. Prompt engineering serves as a critical control mechanism during our research, providing strict guidance that embeds ethical considerations directly into the model's response generation process. However, technological safeguards alone are insufficient for work of this historical significance. Our implementation of human-in-the-loop review has demonstrated that expert validation by historians, linguists, and domain specialists remains critical for sensitive domains such as the Holocaust. Through iterative feedback loops and expert oversight, our results from the defined experiments were validated by experts in the field, and edge cases were identified where LLMs required additional guidance. The human-in-loop approach ensures that generated content not only meets technical accuracy standards but also upholds the ethical, emotional, and commemorative dimensions that Holocaust research demands.

Our adoption of comprehensive data protection protocols has ensured the privacy and integrity of sensitive materials throughout the research pipeline. For releasing the data, we followed the guidance given by the partner museums and archives, who are the stewards of these Holocaust testimonies. Our implementation of access controls and compliance with international regulations, such as the GDPR, is needed to protect the dignity of survivors and uphold our institutional agreements. This commitment to data security is not merely a procedural requirement but

a fundamental ethical obligation with respect to the trust placed in researchers handling such personal and historically significant material.

Our experiments demonstrate that these four techniques are not isolated interventions but rather interconnected elements of a coherent ethical framework. When deployed together with domain-specific fine-tuning and RAG with context-enriched prompts offering behavioural guidance, human-in-the-loop review ensuring expert validation, and data protection protocols safeguarding sensitive materials. The proposed integrated methodology builds a pathway forward for researchers seeking to utilise the capabilities of AI while upholding the highest standards of historical accuracy, ethical responsibility, and respect for survivor testimony. Also, we assume future research will continue to refine these techniques and explore their applicability to other sensitive domains that require similar ethical rigour. However, it is crucial to acknowledge that these safeguards do not eliminate all potential risks. Contemporary LLMs, regardless of the fine-tuning, remain limited in their capacity to replicate the contextual depth, interpretive judgement, and ethical discernment that human experts bring to historically sensitive domains. Therefore, stakeholders who are deploying these applications must approach generated content with appropriate epistemological caution, engaging in critical evaluation of responses within broader and interpretive contexts, and relying on specialised human expertise to validate, contextualise, and determine the factual and ethical acceptability of model outputs. The role of LLMs in Holocaust studies should be understood as a supportive pathway for enhancing research efficiency and accessibility, rather than a supplement to human scholarship and oversight.

8.6 Future of Generative AI

Since GenAI models are designed for automatic text generation, they lack the ability to explain human-like cognitive reasoning. As a result, model outputs need to be critically evaluated before coming to a judgement directly. Consequently, users

of GenAI must be acutely aware of the models' intended purpose and carefully align their application with appropriate settings and objectives. The issues raised by the above section are common to all unstructured sensitive areas, which need additional intervention to process, not only the Holocaust. When engaging with AI-generated materials, particularly in fields that require historical accuracy, users should exercise caution and cross-reference the information with verified sources. Given the inherent limitations of current generative models, including the risk of hallucination, users need to approach AI-generated content with a healthy level of scepticism and awareness of its potential shortcomings. This is especially important when the content involves sensitive topics, such as historical events or mental health, where inaccuracies can have serious consequences.

The outlined limitations of LLMs, such as the absence of comprehensive human understanding, the persistent risk of hallucination, and the risk that generative models can be manipulated to distort history, rightly demand extreme caution whenever GenAI is asked to produce or rephrase historical content on its own. For sensitive, unstructured, or contested topics such as the Holocaust, reliance on fully autonomous generative response is therefore inadvisable and potentially harmful. This does not mean that AI should be rejected totally in Holocaust research and other sensitive domains. With models explicitly designed for pattern recognition rather than free-form generation and following a controlled, human-supervised workflow, AI can deliver scholarly benefits. The key distinction lies in purpose and architecture: generative AI that creates unacceptable risks in sensitive domains, whereas retrieval-augmented systems, classical deep learning, computer vision, and carefully constrained language models used for indexing, transcription, pattern detection, and knowledge linking can dramatically accelerate and enrich research while preserving historical integrity.

The following subsections focus on the dangers of generative AI to concrete, responsible applications that leading archives, historians, and technologists are already implementing successfully. These examples illustrate how AI can serve

as a powerful supplementary tool under critical human oversight to make the historical record more accessible, interconnected, and analytically fruitful without compromising truth or dignity.

8.6.1 Information Retrieval for Research Purposes

The intersection of humanity and information retrieval lies in the ability to use technology to store and retrieve data and preserve, interpret, and humanise the stories embedded within it. This is significant in the context of testimonies of war survivors, where eyewitness or perpetrator narratives contain deeply personal and historical insights that must be handled with accuracy, sensitivity, and ethical responsibility. However, AI is providing powerful tools for learning purposes by processing and analysing large historical records, making content searchable, and providing new avenues for discovery, despite limitations in the underlying database or the design of the algorithms.

It is possible to use AI to uncover "structures, patterns, and trends that are not discernible when the focus remains on just a handful of close readings of individual texts" (Presner, 2024). Consequently, by facilitating the accessibility of historical materials, AI systems are increasingly being used to advance Holocaust studies in a variety of academic fields. For example, data about Holocaust victims was extracted from documents kept in the Arolsen Archives using AI, and it assisted in indexing documents that were specifically difficult and time-consuming for humans to process, like prisoner and transfer lists (B. C. G. Lee, 2019; Archives, 2022). Moreover, deep learning has been employed to analyse sentiment in Holocaust testimonies to better understand the context of family memories preserved in historical archives (Blanke, Bryant, and Hedges, 2020), and similarly, AI can be leveraged to examine digitally created materials such as social media content addressing Holocaust memory to assess how authoritarian regimes instrumentalise the past (Makhortykh, Lyebyedyev, and Kravtsov, 2021). These approaches realise the advanced capacities of AI for recognising patterns in data and can be used to

generate new insights about existing historical materials.

8.6.2 Digital Preservation and Knowledge Linking

Generative AI offers the potential to transform the digital preservation and interconnection of Holocaust material. Many existing archives contain fragile, incomplete, and corrupt records, such as letters, diaries, lists of transfers, and documentation from the camps, which are at risk of being lost over time. Through advanced image restoration, handwriting recognition, and text reconstruction, generative AI could help restore these materials to a legible and analysable form with the support of humans without altering their historical integrity. For example, AI-driven tools can enhance faded ink, predict missing words based on reliable linguistic patterns, and transcribe multilingual handwritten notes into searchable digital text. Making them accessible ensures the longevity of the documents but also makes them accessible to broader audiences, including educators, researchers, and descendants of survivors. Beyond preservation, generative AI serves as a powerful tool for knowledge linking in fragmented or dispersed pieces of historical information in archives and languages. By identifying relationships among names, places, events, and dates, AI can reconstruct contextual and linguistic networks that were previously difficult to trace manually. For example, it associates a name found in one transport list with a letter from another archive or links personal testimonies with official records, enabling a more holistic understanding of individual and collective experiences during the Holocaust.

Such AI-enabled connections also contribute to richer historical narratives and data-driven research. By automatically generating metadata, clustering similar documents, and mapping relationships, AI systems could help historians visualise complex historical phenomena such as migration routes, family separations, and administrative structures of persecution. Ultimately, generative AI-supported digital preservation and knowledge linking not only safeguard material traces of the Holocaust but also deepen our capacity to interpret, connect, and remember

them responsibly for future generations.

8.6.3 Stakeholder Engagement and AI Transparency

In the digital era, our connection with knowledge entails three stages. First, we must efficiently find information: generative AI transforms how we search, discover, and access knowledge, particularly for educational purposes, where learners need relevant, contextualised information delivered quickly. Second, we must preserve information: digital tools enable us to archive historical documents, link fragmented and distributed knowledge across platforms, especially about historical events, and keep the data accessible for future generations. Finally, we have to ensure information remains truthful: AI systems become gatekeepers of information, and we need to take steps to prevent distortion, misinformation, and the rewriting of history. Without an appropriate ethical framework, transparency, and accountability in AI design, the information we retrieve and preserve risks becoming corrupted. This three-stage framework reveals that technological capability alone is insufficient; we must embed safeguarding tools into every layer of our AI-powered information ecosystem. Precisely because these safeguards are not purely technical, their implementation requires a coordinated effort from key stakeholders. In that case, we have to consider the stakeholders who engage with the process, such as policy makers, AI platform developers, and those working for archives, memorials and museums.

For policymakers: It is the duty to integrate ethical AI principles into policymaking processes in the future to guide the responsible development and deployment of AI technologies. Further, as the stakeholders in this ecosystem, policymakers have to instruct AI systems to uphold the historical accuracy, dignity, and integrity in practical scenarios while preserving principles such as fairness, transparency, accountability, and respect for human rights. However, to balance human-AI collaboration, interdisciplinary research projects can be conducted in the future by involving historians, computer scientists, ethical experts, and social

scientists in developing AI tools for monitoring and countering Holocaust denial and distortion online. Due to the expansion of AI, the urge to have regulatory frameworks that address the spread of prejudice and disinformation through AI-powered platforms has increased. As a result, policymakers should encourage these digital platforms to use transparent AI-powered content moderation tools that are automatically capable of detecting and flagging sensitive information related to the Holocaust or contexts that promote Holocaust denial and distortion. Moreover, these guidelines created by policymakers should be exchanged among international audiences by cooperating with joint initiatives among governments and other authoritative organisations and services to develop global cooperation and coordinate responses to AI-amplified Holocaust distortion.

When developing educational systems, comprehensive programmes must equip learners with digital literacy and AI competency to navigate disinformation, prejudice, and hate speech effectively. In this effort, educators play a critical role in alerting learners to AI-generated disinformation and misinformation about historical events such as the Holocaust while teaching them to identify and recognise distorted content. Furthermore, in the future, computer science and ICT curricula in the education sector should integrate critical thinking modules that examine how computing technologies affect society, particularly through the amplification of disinformation and hate speech. When creating policies for the education sector, learners need to be provided with fundamental knowledge regarding the Holocaust while extending their analytical skills to evaluate sources and evidence critically. Consequently, Holocaust museums, archives and memorial institutions need to support the development of AI-focused educational programmes and the integration of digital literacy components into their existing curricula. This approach provides learners with the historical understanding and the technical skills necessary to identify misrepresentation in the digital age.

For developers: AI developers and technical experts must adopt a holistic approach combined with the international human rights standards when developing

digital tools for sensitive domains such as the Holocaust. This foundation must guide all content moderation and curation policies, whether enforced algorithmically or by human moderators. To be effective, this process necessitates moderators with expertise in local languages and cultural contexts. Transparency and verifiable content are the foundation of this trustworthy ecosystem. Developers have to maintain openness about their ecosystems' operations, implementing understandable and auditable policies alongside standard performance metrics. To directly support content authenticity, they could integrate technical solutions such as cryptographically signed metadata (e.g. C2PA Content Credentials) for generative AI media. This enables distributors and audiences to identify AI-generated content, a capability that must be supported by future forensic expertise, including identification icons, interstitial warnings, and robust watermarking compliant with global regulations.

Additionally, the integrity of AI tools is fundamentally dependent on high-quality, inclusive data. Therefore, continuous evaluation of training data quality is essential, with established procedures to ensure accurate and representative data collection. Crucially, this technical work cannot occur in isolation. It requires active partnerships with Holocaust survivors, descendants, and Jewish communities to ensure their voices directly shape policymaking. Furthermore, developers must consistently consult with a wide range of experts, community groups, and organisations to understand their concerns and experiences. Beyond the development phase, AI developers must provide accessible information and tools with multilingual support that allow users to make informed decisions about digital services. This will be accountable to all relevant stakeholders, such as users, the public, advertisers, and regulatory bodies, by implementing terms of service and content policies. Finally, in the future, it is important to develop forensic expertise and technical solutions such as identification icons, interstitial warnings, and watermarking that are compliant with different regulatory approaches to help the broader public and media easily identify fabricated Holocaust content.

For archives, museums and collections: Continued digitisation of Holocaust historical collections remains essential for expanding the data available to train AI systems and enhance their performance. As the domain experts, clear guidelines governing which information AI systems should have access to when retrieving and generating outputs should be stated, and these guidelines should address critical legal and privacy considerations, including copyright protections and the privacy rights of victims and survivors.

Archives, museums, and memorials should strategically adopt AI systems that provide enhanced access to Holocaust information. Furthermore, they have to train researchers on these systems and educate users about how the technology influences search results and information discovery. Equally important is the role these institutions play in guiding AI developers, where they could guide the developers to understand the historical sensitivities, contextual meaning, and ethical risks inherent in incorporating Holocaust databases into AI models. This collaborative approach, where institutions adopt AI technologies and provide guidelines for the responsible use of Holocaust-related information, creates a framework combined with historical integrity while embracing technological innovation in Holocaust education and remembrance.

As researchers, it's our duty to critically assess the possibilities and risks of AI systems, focusing specifically on their behaviour when handling Holocaust information and how various stakeholders utilise them. Such research is critical for assessing the opportunities and risks AI systems pose to Holocaust education and remembrance. Building on this foundation, researchers have to facilitate standardised monitoring and evaluation frameworks that assess how different AI applications handle Holocaust-related resources. These frameworks must include clear performance criteria to improve AI system accuracy, sensitivity, and reliability.

Furthermore, developing technical solutions is essential for identifying Holocaust distortion in the digital ecosystem. Researchers have to develop tools such as identification icons, interstitial warnings, and digital watermarking according to

the different regulatory approaches across jurisdictions. These solutions have to be accessible and easily interpretable, enabling the broader public and media professionals to identify fabricated/manipulated Holocaust content readily.

8.6.4 Countering Historical Denial and Memory Distortion at Scale

The rapid proliferation of GenAI presents a contradictory challenge for Holocaust memory and education. While these technologies offer unprecedented opportunities for preservation and accessibility, simultaneously they enable the systematic spread of historical distortion and denial at scale. As discussed above, GenAI models are prone to "hallucinating" events, personalities and even historical events due to insufficient access to data. ChatGPT and Google's Bard have both produced content detailing Holocaust-related events which never existed, raising critical concerns about misinformation propagation. AI developers may inadvertently train generative AI tools on data from Holocaust denial websites due to a lack of supervision, guidance and moderation. AI has been documented to enable bad actors to distort Holocaust-related content, creating fabricated testimonies and even altering historical records (Rosenthal, 2025).

The scale of AI-generated misinformation represents a qualitative shift in the threat landscape. Deepfake content created using GenAI is compelling for young people, who may encounter it on social media platforms (Nightingale, Wade, and Watson, 2017). This technological shift is particularly concerning for online Holocaust education; with young people relying on GenAI increasingly, exposure to distorted narratives becomes both more likely and more insidious. By utilising NLP techniques and machine learning algorithms, AI can analyse the language, sentiment, and structure of social media posts and news articles to identify potentially misleading and false content. Yet this approach faces inherent limitations because of the potential biases encoded into AI algorithms, which tend to exacerbate existing inequalities, reinforce harmful stereotypes and blur the distinction between

disinformation and legitimate speech. Museums and archival institutions are emerging as critical stakeholders in this digital ecosystem. The Auschwitz museum discovered Facebook pages producing Holocaust victim biographies with fictional information and AI-generated photographs, with museum officials warning that "producing AI images of real people, or what is even more troubling, producing false identities of victims, is certainly problematic and brings a tainted nature to the memory of those who actually sacrificed their lives at Auschwitz" (The Times of Israel Staff, 2025). This institutional response models an essential counterbalance: the systematic curation and amplification of authentic materials to saturate the information environment with verified content.

The trajectory toward scalable counter-denial requires shifting from reactive mitigation to proactive truth amplification. The mass digitisation of records and their integration into AI systems can go some way towards protecting the record of the Holocaust from erasure or distortion. Yet AI can only draw from the information it is trained on, meaning that if AI models have only limited narratives, they will produce flawed outputs or reproduce the same well-known stories over and over again, amplifying some narratives while eroding the breadth and depth of the history of the Holocaust. A future approach must prioritise: (1) enrichment of training datasets with verified, multilingual archival materials; (2) development of domain-specific detection models sensitive to coded antisemitic discourse; (3) integration of AI-driven content moderation with expert review; and (4) public digital literacy initiatives enabling users to critically evaluate AI-generated content. Most fundamentally, future approaches must protect the facts and encourage critical thinking while taking advantage of the new opportunities AI offers to strengthen understanding of and education about the Holocaust and make such education and information available to all.

8.7 Chapter summary

This chapter examined the ethical dimensions of applying LLMs to sensitive historical corpora, using Holocaust survivor testimonies. It addressed the fourth and final research question of this thesis: what ethical considerations arise when deploying LLMs to process sensitive oral historical narratives, and how should these shape the design of NLP systems in digital humanities contexts? The chapter opened by establishing that the integration of LLMs into oral history practice is not a neutral methodological development but one that creates fundamental tensions with the field's foundational ethical commitments. The traditional pillars of oral history (informed consent, relational stewardship, and contextual fidelity) were shown to be structurally incompatible with how LLMs operate because LLMs act as decontextualising, probabilistic engines that process human narrative as tokenised data without regard for its interpersonal, cultural, or commemorative dimensions.

Building on this foundation, the chapter identified four categories of ethical risk that emerge specifically from the technical characteristics of LLMs. First, the consent and ownership framework developed for human-conducted oral history research is rendered inadequate by the opacity, open-endedness, and irreversibility of LLM training, where the narrator's voice, once fed into model parameters, cannot be meaningfully withdrawn. Second, the extraction and repurposing of culturally specific narratives by LLMs developed and owned by powerful institutions risks reproducing neo-colonial dynamics of data extraction, where communities whose experiences form the basis of training data receive neither recognition nor benefit. Third, the statistical nature of LLMs poses a direct threat to historical accuracy through hallucination, overgeneralisation, and the systematic marginalisation of testimonies. Fourth, the re-identification risk inherent in LLM pattern recognition means that standard anonymisation techniques are insufficient to protect the privacy and dignity of survivors and their families.

This chapter contributes a novel framework to mitigate these ethical risks. Rather than treating ethics as an abstract commentary on AI risk, this chapter

demonstrates how ethical obligations can be operationalised at every stage of an NLP pipeline. The four mitigation strategies — domain-specific fine-tuning augmented by retrieval-augmented generation, context-enriched prompt engineering, human-in-the-loop expert validation, and robust data protection protocols — were not adopted in isolation but deployed as an integrated framework in which each component reinforces the others.

The chapter concluded by looking forward, identifying responsible applications of AI in Holocaust research, such as information retrieval, digital preservation, knowledge linking, and counter-denial at scale. Furthermore, this chapter outlines the responsibilities of policymakers, developers, archival institutions, and researchers in ensuring that these applications are deployed with transparency, accountability, and genuine respect for the communities whose histories they engage. The high-level argument is that the role of AI in Holocaust studies, and in sensitive digital humanities research more broadly, should be understood as a carefully bounded, human-supervised tool for enhancing research access and efficiency, not as a substitute for the interpretive judgement, ethical discernment, and commemorative responsibility that only human scholars can provide.

Chapter 9

Conclusions

'Never Again': Never again becomes more than a slogan: it's a prayer, a promise, a vow

United States Holocaust Memorial Museum / Elie Wiesel

9.1 Introduction

This thesis set out to address a dual challenge: how to preserve the integrity and accessibility of oral narratives of war survivors for future generations and how to do so in a way that is both computationally principled and ethically responsible. In this research, testimonies related to the Second World War were being selected, and hundreds of thousands of oral testimonies held in archives around the world represent an irreplaceable corpus of human memory. However, according to our observations, the available computational tools could not handle the linguistic complexity or moral importance of these narratives. In response, this thesis developed and validated an end-to-end NLP framework for Holocaust survivor testimonies, showing that computational access to these narratives can be achieved while preserving historical accuracy and ethical responsibility. Four research questions structured the investigation, each targeting a distinct component of that overarching problem. The following sections revisit each question in turn, summarise the contributions made in answering it, acknowledge the limitations of the current work, and outline

the directions that future research should take.

9.2 Summary of Contributions

This thesis makes four original contributions to the fields of natural language processing and digital humanities, summarised below:

1. **A domain-aware knowledge extraction and representation framework.** An end-to-end pipeline comprising domain-adapted transformer models for named entity recognition, a toponym disambiguation system, an LLM-based relationship extraction approach, and a knowledge graph construction pipeline that transforms unstructured oral testimony transcripts into structured, queryable RDF knowledge representations.

2. **A curated and annotated Holocaust oral testimony corpus.** The first manually annotated gold corpus of Holocaust survivor testimonies designed for NLP research, comprising 200 testimonies annotated with a domain-specific schema covering named entities, relationships, and contextual markers. A pseudo-labelling methodology is also introduced to partially automate annotation across the remaining corpus, with transferable rule-based and prompt-engineering components adaptable to other historical or cultural domains.

3. **DsQoLA - a lightweight domain-specific adapter for RAG pipelines.** A parameter-efficient adapter architecture that improves retrieval accuracy over domain-specific historical corpora by learning a linear transformation of the query embedding space, without retraining the underlying model or recomputing the document index.

4. **An ethical framework for LLM processing of sensitive historical narratives.** A domain-specific, integrated mitigation framework that identifies four categories of ethical risk arising from LLM deployment on Holocaust

testimony and proposes four interconnected mitigation strategies validated throughout the research pipeline.

The following subsections examine how these contributions collectively address the four research questions that structured the investigation.

RQ1: How can NLP techniques, including domain-aware entity recognition, prompt-based relationship extraction, and knowledge graph construction, be combined into a coherent framework for extracting and representing structured knowledge from Holocaust oral narratives?

Chapters 4, 5, and 6 together constitute an end-to-end framework for transforming raw, unstructured oral testimony transcripts into structured, queryable, and analytically rich knowledge representations. Chapter 4 presented a domain-aware NER system built on transformer models further pre-trained on Holocaust-specific testimony corpora using a masked language modelling objective (Devlin et al., 2019), producing three adapted models (HoloBERT, HoloRoBERTa, and HoloXLNet), of which HoloBERT achieved the lowest perplexity (3.1259). Fine-tuned on the labelled NER corpus, these models substantially outperformed multilingual baselines across Holocaust-specific entity categories. Furthermore, chapter 4 introduced a methodology for toponym disambiguation in Holocaust testimonies, addressing the historically specific challenge that place names carry multiple geopolitical, administrative, and emotional meanings within the same corpus. A few-shot chain-of-thought prompting approach enriched with a structured Holocaust knowledge base outperformed zero-shot and RAG-based alternatives (P. Lewis, Perez, Piktus, Petroni, Karpukhin, Goyal, Küttler, M. Lewis, Yih, Rocktäschel, et al., 2020a) on domain-specific geospatial categories, including GHETTO (+0.19 F1) and LOC (+0.08 F1). Chapter 5 advanced the framework to relationship extraction, introducing an LLM-based approach using few-shot prompt engineering to identify structured relational triplets from testimony narratives. Rule-based and argument-mining approaches were evaluated and found insufficient for the unstructured, emotionally charged discourse of survivor testimony. The few-shot

prompting approach, structured around three domain-specific relationship categories — biographical, career-based, and Holocaust event-based — achieved F1 scores of 96.4%, 83.78%, and 85.4% respectively against the human-annotated gold standard. Chapter 6 integrated these extraction outputs into a comprehensive knowledge graph pipeline, formalised in RDF and stored in GraphDB, introducing domain-specific preprocessing components including LLM-based coreference resolution, relationship normalisation, and URI-based entity linking to external Holocaust archives. Applied to 98 testimonies from the Centropa archive, the resulting graph contains 5,613 nodes and 6,146 edges. Graph analytics revealed 47 thematic communities with a modularity score of 0.7923, a power-law degree distribution centred on historically significant hubs, and three dominant subgraph patterns corresponding to deportation sequences, family separation, and multi-generational residence – structural findings unattainable through close reading or conventional archival analysis alone.

RQ2: What are the linguistic and structural challenges of Holocaust oral testimonies as unstructured data, and what corpus design and annotation strategies are required to make them computationally processable? The linguistic analysis presented in Chapter 2 and the domain-specific markers identified in Holocaust testimonies in Chapter 3 revealed most of the linguistic properties which exist in Holocaust testimonies are not found in any existing general-purpose NLP training corpus. These include narrative structures shaped by code-switching across Yiddish, German, Polish, Hebrew, and other European languages within a single testimony; historically ambiguous spatio-temporal references; domain-specific terminology covering concentration camps, ghettos, and military units that falls outside standard NER taxonomies; and paralinguistic features including pauses, repetitions, and emotional digressions. No annotated resource existed for this domain at the outset of the research, making the construction of a gold standard the necessary first step. A domain-specific annotation schema covering named entities, relationships, and contextual markers

was developed in collaboration with domain historians and validated through a systematic comparison of three annotation approaches: manual expert annotation, semi-automated annotation using SpaCy, and LLM-based annotation using prompt engineering. The resulting gold corpus of 200 manually annotated testimonies constitutes the first publicly available annotated resource of its kind and underpins all subsequent technical contributions in the thesis. As a secondary contribution, the chapter introduced a pseudo-labelling methodology that partially automates the annotation process across the remaining 2,800 testimonies, reducing both the financial and the labour cost required to construct labelled corpora at scale. Rather than treating SpaCy and LLM-based annotation as independent tools, the framework establishes a principled selection process by validating against the gold standard that determines which method is most appropriate for each annotation task. The domain-specific rules and prompt strategies developed within this process are designed for transferability: the rule-based components can be adapted for annotation in different historical or cultural corpora by substituting domain-specific vocabulary, and the prompt engineering strategies can be reused across domains with minimal modification, requiring only the replacement of Holocaust-specific terminology with terminology relevant to the target corpus.

RQ3: How can a lightweight, parameter-efficient adapter architecture improve retrieval accuracy in RAG pipelines for domain-specific historical corpora without requiring large-scale model retraining?

When a retrieval system fails to identify a contextually relevant passage, the generative model downstream is likely to hallucinate or produce historically incomplete responses (Z. Ji et al., 2023). In a domain such as Holocaust testimony, where factual precision and contextual sensitivity are ethical obligations rather than merely performance metrics, this retrieval bottleneck is particularly consequential. General-purpose embedding models trained on web-based text fail to capture the multilingual vocabulary, domain-specific terminology, and cultural context of testimony corpora. Moreover, fully retraining these models

is neither computationally practical for most digital humanities environments nor sustainable as archive collections grow and are updated. Chapter 7 addressed this through DsQoLA, a parameter-efficient adapter architecture that improves retrieval accuracy by learning a lightweight linear transformation of the query embedding space, aligning query representations with the distributional characteristics of testimony document embeddings, without modifying the underlying base model or recomputing the document index (Houlsby et al., 2019). Trained on 74,343 query-passage pairs synthesised from 1,783 testimonies using a hybrid InfoNCE and Triplet Margin Loss with hard negatives from a contrastively distant financial news corpus, DsQoLA produced consistent improvements in Hit@K and MRR across four base embedding models. The adapter-enhanced multi-qa-mpnet-base-dot-v1 model achieved the strongest absolute scores (Hit@10: 0.4863; MRR: 0.2903), while stsb-roberta-base showed the largest relative gains (+63.2% Hit@1; +59.6% MRR). The finding that DsQoLA is most effective at compensating for weaker base models has particular practical significance for resource-constrained digital humanities environments that lack access to high-performance computing infrastructure.

RQ4: ***What ethical considerations arise when deploying large language models to process sensitive oral historical narratives, and how should these shape the design of NLP systems in digital humanities contexts?***

Chapter 8 highlighted the issues between the technical design of LLMs and the ethical principles of oral history. LLMs process text through statistical prediction and standardised tokenisation (Bommasani et al., 2021), whereas oral history prioritises informed consent, contextual integrity, and responsible stewardship of personal narratives. According to our analysis, four categories of domain-specific ethical risk were identified: existing consent frameworks do not account for AI processing of archival material; powerful institutions risk neo-colonial data extraction from culturally sensitive collections; training data biases may produce hallucination or historical misrepresentation; and standard anonymisation techniques fail to prevent re-identification through LLM pattern inference. The novel

contribution of this chapter is not the identification of these risks in isolation but the proposal of a domain-specific, integrated mitigation framework that addresses all four concurrently within a sensitive historical NLP context, validated throughout the research pipeline rather than proposed purely theoretically.

9.3 Limitations

The limitations of this thesis are considered in direct relation to each of the four research questions, reflecting the boundaries of what the current work can and cannot claim.

Knowledge Graph Coverage and Computational Resource Constraints. (RQ1)

The knowledge graph construction pipeline was applied to a sample of 98 testimonies drawn exclusively from the Centropa archive, owing to the hardware and computational resource constraints encountered during the research. These structural findings are analytically significant and validate the pipeline, but they are not representative of all 3,000 testimonies or of Holocaust survivor experience more broadly. The generalisability of these findings to testimonies from other archives, geographical regions, and linguistic backgrounds remains an open empirical question that the current work does not resolve. Nevertheless, the pipeline itself is not constrained to this sample. The knowledge graph construction methodology, including the RDF formalisation, entity resolution, relationship normalisation, and URI-based linking components, was designed to be scalable and is directly applicable to the full testimony corpus given sufficient computational infrastructure. Beyond the Holocaust domain, the same pipeline could be adapted for knowledge graph construction from other oral historical corpora, providing a reusable technical foundation for digital humanities projects working with testimony collections at scale.

Corpus Scale, Linguistic Diversity, and Annotation Scope. (RQ2)

The gold-standard corpus of 200 manually annotated testimonies was sufficient to enable comparative evaluation and to produce meaningful experimental results within the scope of this research. However, the dataset is not large enough to support the standalone training of NLP models in the manner typical of traditional large-scale supervised learning approaches. Consequently, the models and annotation comparisons presented in this thesis should be understood as demonstrating a domain-adapted methodological framework rather than production-ready systems. Furthermore, the testimonies included in the corpus were selected from English-language transcripts and therefore do not completely represent the linguistic and geographical diversity of Holocaust survivor experiences, many of which were originally delivered in languages such as Yiddish, Polish, Hebrew, German, and other European languages. Finally, the annotation schema developed in this research focuses on textual features and does not capture paralinguistic aspects of oral testimony, such as pauses, tonal variation, and non-verbal emotional expression. These elements often carry significant meaning in oral history discourse but remain outside the scope of current text-based NLP methods.

Retrieval Performance and Training Data Availability. (RQ3)

Although DsQoLA produced consistent and meaningful improvements in retrieval accuracy across all evaluated configurations, the absolute scores achieved by the best-performing configuration — Hit@10: 0.4863 and MRR: 0.2903 which indicate that retrieval performance remains at a level that would be insufficient for deployment in a production-grade archive search or question-answering system. This limitation depends on two factors. The first is intrinsic to the task: oral testimony text is linguistically irregular, unstructured, and contextually rich in ways that make passage-level retrieval substantially harder than retrieval over well-structured documentary or encyclopaedic text. The second source is a data constraint: the absence of large-scale, human-annotated relevance judgements for Holocaust testimony retrieval meant that the query-passage training pairs used to train DsQoLA were synthetically generated rather than human-validated, which

limited both the quality and the diversity of the supervision signal available to the adapter. Addressing this limitation would require the construction of a dedicated retrieval evaluation benchmark for Holocaust testimony, with human-annotated relevance judgements produced in collaboration with domain experts.

Ethical Framework Validation and Hallucination Risk. (RQ4)

The ethical framework proposed in Chapter 8 represents a theoretically grounded and domain-specific contribution to responsible NLP system design. However, it was developed within the scope of this research as a conceptual framework rather than being validated through formal empirical methods such as structured stakeholder consultations, independent user studies, or pilot deployment within a live archival environment. The framework reflects considered judgements informed by engagement with domain historians, archivists, and the relevant literature on AI ethics. Nevertheless, its practical effectiveness under real-world conditions remains to be demonstrated through future collaborative validation with Holocaust memorial institutions.

A related limitation concerns the potential for hallucination and historical inaccuracy in outputs generated by large language models. Despite the human-in-the-loop validation protocols incorporated throughout the proposed research pipeline, the current architecture cannot completely eliminate the possibility that a model may generate outputs that are factually incorrect, historically misleading, or contextually inappropriate in relation to survivor testimony. Further mitigation will require advances in domain-specific grounding techniques, more rigorous output verification mechanisms, and sustained collaboration with historians capable of identifying subtle historical inaccuracies that automated evaluation methods may fail to detect.

These limitations define the scope within which the contributions of this thesis should be understood, and each identifies a concrete direction for the future research outlined in the following section.

9.4 Future Directions

From this research, several challenges were identified that need to be addressed in the future. This presents opportunities for further exploration within the digital humanities that extend beyond the scope of the current PhD thesis.

- **The multilingual nature and the code-mixing nature of the Holocaust testimonies.**

  Code-mixing is a common scenario in Holocaust testimonies where words, phrases and grammatical structures from different languages blend within a single discourse. Holocaust survivors came from diverse linguistic and cultural backgrounds from multiple European territories, each with its own language. As a result, survivors frequently switch between languages when describing locations, events, and cultural incidents that happened during the Holocaust. Moreover, when survivors recount their memories, they switch from a single language to their native language, specifically when describing deeply distressing or traumatic events. This transformation often occurs because extreme emotions, such as fear, pain, or sorrow, can trigger a natural return to the language most closely tied to their identity and early experiences. However, state-of-the-art models struggle to recognise code-mixing due to the limited availability of high-quality annotated datasets compared to monolingual data.

- **Standardisation of the Holocaust Testimonies:**

  Standardisation of Holocaust testimonies is a complex task when digitalising as a historical resource. Since different organisations use different objectives and policies to collect and preserve, bringing all of them to a common ground is difficult. Each organisation operates with its mission, target audience, and resources, leading to differences in how testimonies are recorded, transcribed, archived, and shared. Some organisations may focus on preserving raw, unedited testimonies to maintain authenticity, while others make it more accessible for research and educational purposes. One of the primary challenges in annotation

is the lack of a standardised dataset format, which makes it difficult to convert Holocaust testimonies into forms suitable for applying NLP techniques.

- **Compatibility of spoken language with LLMs.**

  The complex linguistic patterns and noise in oral language make it more challenging to analyse than written language. Informal speech patterns, such as interruptions and overlapping dialogue, are common in conversations. Oral language is often unstructured, consisting of fragmented sentences, pauses, repetitions, and interruptions. Unlike written text, which follows grammatical rules and structured syntax, spoken language is context-dependent, making it difficult for NLP models to parse and analyse accurately. When considering the Holocaust as a historical event, spoken language evolves, incorporating new slang, idioms, and expressions. In testimonies, survivors use terms that were common in wartime ghettos, concentration camps, or resistance movements, many of which do not exist in modern vocabulary. State-of-the-art NLP models trained on general datasets may struggle to understand domain-specific, historically and culturally specific terminology, limiting their applicability in Holocaust testimony analysis.

- **Generalisation to Other Sensitive Oral History Domains**

  Although this research was developed and validated specifically in the context of Holocaust survivor testimony, the methodological framework it proposes is not inherently restricted to this domain. It was designed to address challenges that recur across a wide range of oral historical corpora involving traumatic, multilingual, and culturally specific narrative content, and its constituent methods are transferable to any domain that shares these fundamental characteristics. Oral history archives documenting other genocides, such as the Rwandan genocide, the Cambodian Khmer Rouge period, and the Bosnian conflict, face structurally similar computational challenges: unstructured, trauma-induced narrative discourse; domain-specific historical and cultural vocabulary; code-switching across regional languages; and different ethical obligations to handle

survivor accounts with dignity and contextual fidelity. Furthermore, the conflict-zone survivor accounts from more recent crises, including those emerging from Syria, Ukraine, and Myanmar, present additional challenges of temporality and incompleteness, where testimony collection is ongoing and the historical record is still being formed. Adapting and validating the ethical framework across these domains represents both a significant research opportunity and, given the urgency of preservation before living witnesses are lost, a moral imperative. Achieving this will require sustained interdisciplinary collaboration between NLP researchers, oral historians, archivists and ethicists.

9.5 Reflection

This thesis began from a simple but urgent motivation of preserving the memories of a generation who lived through the Holocaust. Within the coming decades, the living memory sustained by survivors will give way to a post-testimonial era in which their voices will be accessible only through the archives they leave behind. Given the scale of the collections and the urgency of preservation, computational engagement is not merely optional but necessary. Accordingly, this research does not ask whether technology should be applied to these archives but rather how it can be applied responsibly and under what ethical conditions to enhance accessibility while safeguarding the integrity of the material.

The framework presented in this thesis offers one answer to that question: build from the ground up for the domain; treat ethical obligations as design constraints rather than compliance requirements; preserve the provenance and contextual integrity of every extracted fact; and maintain human expertise at the centre of every critical decision in the pipeline. These principles are not unique to Holocaust testimony. They apply wherever computational systems are asked to mediate access to narratives of human suffering, cultural memory, and historical truth. The survivors whose testimonies form the basis of this research bore witness

so that future generations would know what happened and why it must never be allowed to happen again. The responsibility of those who work with their testimonies computationally is to ensure that technology serves that purpose by making their words more accessible, more interconnected, and more durable without diminishing the humanity, complexity, and moral weight that give those words their lasting significance. This thesis represents one step in that direction.

Appendix A

Background and Related Work

A.1 Digital Archives and Institutional Repositories for Holocaust Survivor Testimonies

The table below lists sources of Holocaust testimonies. Some are publicly accessible, while others require authorisation from the respective institutions that hold them.

Table A.1: Institutions and Testimony Archives

Institute/Archive name	**Number of testimonies**	**Web link**
USC Shoah Foundation's Visual History Archive	55,000+	sfi.usc.edu
Yad Vashem's Testimony Collections	13,000+	yadvashem
United States Holocaust Memorial Museum (USHMM)	15,000+	ushmm
Fortunoff Video Archive for Holocaust Testimonies	4,400+	fortunoff
Wiener Holocaust Library	1,000+	wiener
Centropa	1,200+	centropa
Montreal Holocaust Museum	800+	museeholo
Museum of Jewish Heritage	4,000+	mjhnyc
Austrian Oral History Collection	1,200	oeaw
NIOD Oral History Collection	1,200	niod
Voices from Ravensbrück	200	ravensbrueck
Bergen-Belsen Memorial Oral History Collection	200	bergen
Ghetto Fighters' House Archive	1200	gfh

Inter-Annotator Agreement

Automatic Extraction of historical information from the Holocaust testimonies

#how do testimonies work?

Holocausts and genocides always bring heart-breaking memories about the 19th century. The people persecuted by the tragedy of the Holocaust are considered as the survivors. The survivors are the actual imprints that claim the horribleness of the Holocaust. Testimonies convey to bear witness by telling what survivors saw or what survivors know, intending to make the audience knowledgeable. Survivors testify about their memories and experiences in different forms of documents such as oral history interviews, transcribed interviews, memoirs, diaries, and letters.

To a large extent and for decades, the Holocaust testimonies have been scattered, endangered and virtually inaccessible. With the advancements in digital humanities, most of these historical documents have been digitised. This gives the opportunity for the computational methods to process them and extract biographical information automatically which is significantly faster than manual processing. The primary goal of this study is to develop and apply state-of-the-art NLP techniques combined with deep learning techniques that could be used to help identify the connections between these Holocaust testimonies. This will provide Holocaust researchers with a tool to explore the testimonial data more efficiently so that new insights can be discovered about this historical atrocity.

Artificial intelligence research is advancing fast in the humanities. However, a major current limitation is the lack of relevant training data that allows the digital humanities to include more advanced methods such as deep learning in their practices. From this research, we have crawled and collected digitally available Holocaust testimonies. Our aim is to bring up centralised knowledge model to identify relationships exists on Holocaust testimonies.

- Data collection and dataset

Nature of the dataset

The collected dataset consists on … testimonies crawled from different holocaust museum websites. In further, Holocaust testimony elaborates on personal experience or memory in a more descriptive nature, explaining before, and after a particular incident. Or either, it takes the form of questions and answers between the interviewer and the survivor. However, the holocaust testimonies consist of lots of noise because it is documented in verbal/oral language.

- Levels of annotation

  The annotation will be carryout in 3 levels. More information about the two levels of the annotation is given below.

- **First Level**
  Annotating the testimony whether is it their own experience (1st person) or explaining it on behalf of another (3rd person). Holocaust survivors explain their memories or some of their close person have experience about the Holocaust. From the first level annotation what mainly focus is to identify wheather the testimony is about a his own self or told on behalf of other person.

- **Second Level -** Annotating the testimony based on the different tags to identify the relationships that exist in individual testimonies such as concentration camps, ships, ghetto, etc. Following given below are the tags which needed to be annotated. Following given below are the tags and relationships that needed to consider in the process of annotation.

- **Third Level -** Annotation based on the activities the survivors have participated in during the Holocaust

Named Entity Tags:
Ghetto, Ships, Concentration camp, City, Spouse, River, Forest, Mountain, Military rank, Street, Country, Date, Lawsuit, Ethnicity, Language, Event, Nationality, Location, Organisation, Disease, person, Military Group, Religion, Gathering, Waterbodies, Mass Arrest, Political Affiliation, Profession

Relationships:
taken to, located in, born in (country), born on(date), killed in (place), killed on (place), arrest in, escape from, escaped to, killed by, worked as, hide on, married to, transported to, deported to, arrived on, left to, lost at, Return to, Death of, survived, worked, Employed, Located, Lostin, Died, Died at, Died In

Classification Types:
Mass shooting, Flogging, Beating, Kicking, Drowning, Waterjet tortures, Emasculation, Systematic selection of prisoners, Continuous shooting, Migration, Suicide, Rescue, Migrants, Homosexuals, Kindertransport, Deportations, Expropriation, Antisemitism, Children, Aryanisation

- Process of the annotation

Examples 01

Sentence	Entity Tags	Relationship
Scharfer, and who returned to Poland from Russia in 1946.	Scharfer, Poland, Russia, 1946	Returned to
Jane was taken to Gestapo headquarters on Kilinski Street for interrogation and indescribable torture. The next day the Gestapo handcuffed me with wire and transported me in a boxcar to Gross-Rosen concentration camp	Jane, Gestapo headquarters, Kilinski Street, Gross-Rosen concentration camp	Taken to, transported

Relations:

Taken to
Jane ----------> Gestapo headquarters
Transported to
Jane ----------->Gross-Rosen concentration

- Annotation tools and guidance

The annotation process continues in the UBIAI Platform. This platform is an annotation tool which supports multiple people to annotate the same document. After creating of account in the UBIAI platform, you can log in to you're assigned task and start the annotation according to the different levels of the annotation guideline. The below screenshots elaborate more information about the steps that needed to follow on the UBIAI annotation platform.

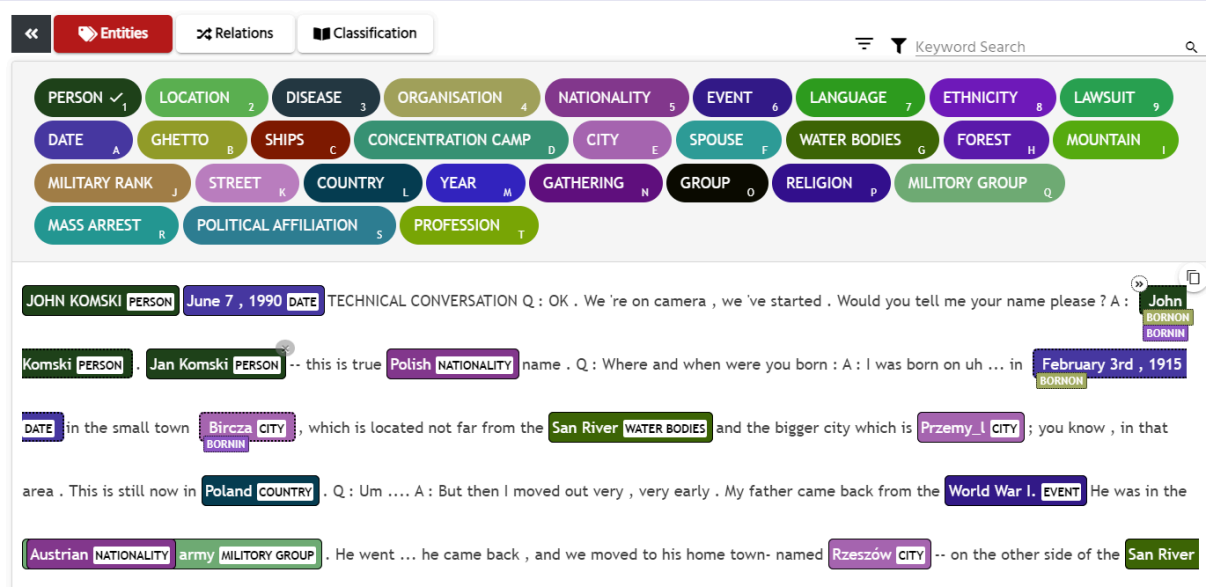

Fig 01: Sample annotation of Holocaust testimonies

- Video guide for using of UBIAI Annotation tool

https://drive.google.com/file/d/1bI93BA88uERsrxAsk9c7-FXLyvBPhDqx/view?usp=sharing

Outcome

After getting the ethical approval to publish it as a dataset, the Following will receive as a credit for the annotators,

- Participants will be able to work with different testimonies of Holocaust survivors from different Holocaust libraries and museums such as Wiener Holocaust Museum, USHMM and Yale University Fortunoff Archive.
- Participants will get the opportunity to be co-authors of the dataset (Publication).
- Participants' names will be added to the final release of the software application under the section of contributors

Further, they will be acknowledged in advance on different occasions in the research.

More Information(We are encouraged to read this section)

As an Expert in History, What You'll Be Doing

- You will read transcribed, publicly available testimonies from Holocaust survivors.
- You will annotate names, places, relationships, and events in the text.
- These annotations will help future research and digital tools, including possible AI applications.

What You Should Know Before You Agree

Please know that we are here for you with open hearts, and thank you very much for your support. If you experience any difficult or overwhelming emotions while reading these testimonies, don't hesitate to reach out to us.

Emotional Risks:

- These testimonies include graphic and distressing content, such as violence, trauma, and loss.
- Reading them may cause emotional reactions such as sadness, distress, anxiety, or emotional numbness.
- You may feel uncomfortable turning deeply personal accounts into data.

We're Here to Support You:

- You'll receive a short training to prepare yourself for the annotation process and explain all the emotional risks.
- **We are offering a limited number (less than 20) of testimonies to annotate, which implies less emotional impact.**
- **As organisers, we conduct weekly meetings to check the emotional distress and other annotation-related problems that occurred during the process.**
- **If any participant feels any emotional distress, we request that they immediately stop the annotation process and either reach out to us or contact the support lines below as soon as possible.**
- **Any participant can withdraw from the annotation if they feel any emotional distress and providing additional support from contacting the NHS link- Psychological talking therapies Or Mental Health support team at Lancaster, in order to get support.**
- You'll also have optional group reflection sessions to talk about your experiences every week.

You Can Say No:

- Participation is voluntary.
- You can withdraw at any time with no penalty or negative impact on your academic standing.
- You can also choose not to annotate specific testimonies if they are too difficult.

After the Project

- We will offer a brief session at the end, but please know that this alone may not be enough to fully process what you've experienced.
- You are encouraged to continue using support services if needed.

References

Abadie, Nathalie et al. (2022). "A benchmark of named entity recognition approaches in historical documents application to 19 th century french directories". In: *International Workshop on Document Analysis Systems*. Springer, pp. 445–460.

Afreen, Juveria, Mahsa Mohaghegh, and Maryam Doborjeh (2025). "Systematic literature review on bias mitigation in generative AI". In: *AI and Ethics* 5.5, pp. 4789–4841.

Agarwal, Dhruv, Mor Naaman, and Aditya Vashistha (2025). "Ai suggestions homogenize writing toward western styles and diminish cultural nuances". In: *Proceedings of the 2025 CHI conference on human factors in computing systems*, pp. 1–21.

Aguilar, Sergio Torres (2022). "Multilingual Named Entity Recognition for Medieval Charters Using Stacked Embeddings and Bert-based Models." In: *Proceedings of the second workshop on language technologies for historical and ancient languages*, pp. 119–128.

Al Nazi, Zabir, Md Rajib Hossain, and Faisal Al Mamun (2025). "Evaluation of open and closed-source LLMs for low-resource language with zero-shot, few-shot, and chain-of-thought prompting". In: *Natural Language Processing Journal* 10, p. 100124.

Anh-Hoang, Dang, Vu Tran, and Le-Minh Nguyen (2025). "Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior". In: *Frontiers in Artificial Intelligence* Volume 8 - 2025. ISSN: 2624-8212. DOI: 10.3389/frai.2025.1622292. URL: https://www.

frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1622292.

Anisuzzaman, D.M. et al. (2025). "Fine-Tuning Large Language Models for Specialized Use Cases". In: *Mayo Clinic Proceedings: Digital Health* 3.1, p. 100184. ISSN: 2949-7612. DOI: https://doi.org/10.1016/j.mcpdig.2024.11.005. URL: https://www.sciencedirect.com/science/article/pii/S2949761224001147.

Anti-Defamation League (Mar. 2025). *Generating Hate: Anti-Jewish and Anti-Israel Bias in Leading Large Language Models*. Press release. Accessed: 14 May 2026. URL: https://www.adl.org/resources/press-release/%20anti-jewish-and-anti-israel-bias-found-leading-ai-models-new-adl-report.

Anuradha, Isuri, Francesca Frontini, et al. (2025). "Proceedings of the First on Natural Language Processing and Language Models for Digital Humanities". In: *Proceedings of the First on Natural Language Processing and Language Models for Digital Humanities*.

Anuradha, Isuri, Le Ha, et al. (2023). "Enhancing named entity recognition for holocaust testimonies through pseudo labelling and transformer-based models". In: *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing*, pp. 85–90.

Anuradha, Isuri, Ruslan Mitkov, et al. (2025). "HoloBERT: Pre-Trained Transformer Model for Historical Narratives". In: *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing-Natural Language Processing in the Generative AI Era*, pp. 105–110.

Anuradha, Isuri, Ruslan Mitkov, Vinita Nahar, et al. (2023). "Evaluating of Large Language Models in Relationship Extraction from Unstructured Data: Empirical Study from Holocaust Testimonies". In: *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pp. 117–123.

Anuradha, Isuri, Tharindu Ranasinghe, et al. (2026). "LiLADH: An Open Retrieval Resource for Digital Humanities Archival Corpora". In: *Submitted to*

the Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM 2026). Under review.

Anuradha, Isuri, Deshan Koshala Sumanathilaka, et al. (2025). "Toponym Resolution: Will prompt engineering change expectations?" In: *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing-Natural Language Processing in the Generative AI Era*, pp. 95–104.

Anuradha, Isuri and Martin Wynne (2026). "Proceedings of the Second Workshop on Holocaust Testimonies as Language Resources (HTRes)@ LREC 2026". In: *Proceedings of the Second Workshop on Holocaust Testimonies as Language Resources (HTRes)@ LREC 2026*.

Anuradha, Isuri, Martin Wynne, et al. (2024). "Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes)@ LREC-COLING 2024". In: *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes)@ LREC-COLING 2024*.

Archives, Arolsen (2022). "Everynamecounts Uses AI to Uncover Information on Victims of Nazi Persecution". In: *Arolsen Archives*.

Asudani, Deepak Suresh, Naresh Kumar Nagwani, and Pradeep Singh (2023). "Impact of word embedding models on text analytics in deep learning environment: a review". In: *Artificial intelligence review* 56.9, pp. 10345–10425.

Atanassova, Iana, Marc Bertin, and Philipp Mayr (2022). *Mining Scientific Papers, Volume II: Knowledge Discovery and Data Exploitation*.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473*.

Bailey, Cassandra et al. (2020). "What are we missing? How language impacts trauma narratives". In: *Journal of Child & Adolescent Trauma* 13.2, pp. 153–161.

Bassignana, Elisa and Barbara Plank (May 2022). "What Do You Mean by Relation Extraction? A Survey on Datasets and Study on Scientific Relation Classification". In: ed. by Samuel Louvan, Andrea Madotto, and Brielen Madureira,

pp. 67–83. DOI: 10.18653/v1/2022.acl-srw.7. URL: https://aclanthology.org/2022.acl-srw.7/.

Bauer, Yehuda (2002). *Rethinking the holocaust*. Yale University Press.

Beard, Martha Rose (2017). "Re-thinking oral history–a study of narrative performance". In: *Rethinking History* 21.4, pp. 529–548.

Beelen, Kaspar et al. (2021). "When time makes sense: A historically-aware approach to targeted sense disambiguation". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2751–2761.

Bergen, Teresa (2019). *Transcribing oral history*. Routledge.

Berry, David M (2012). "Introduction: Understanding the digital humanities". In: *Understanding digital humanities*. Springer, pp. 1–20.

Blanke, Tobias, Michael Bryant, and Mark Hedges (2020). "Understanding memories of the holocaust—A new approach to neural networks in the digital humanities". In: *Digital Scholarship in the Humanities* 35.1, pp. 17–33.

Blouin, Baptiste et al. (2021). "Transferring Modern Named Entity Recognition to the Historical Domain: How to Take the Step?" In: *Workshop on Natural Language Processing for Digital Humanities (NLP4DH)*.

Boder, David P. (Sept. 1946). *Interview with Jürgen Bassfreund, September 20, 1946*. Voices of the Holocaust, Illinois Institute of Technology. Accessed: [Insert Date Here]. Munich, Germany. URL: https://voices.library.iit.edu/interview/bassfreundJ.

Bojić, Ljubiša et al. (2025). "Comparing large Language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm". In: *Scientific reports* 15.1, p. 11477.

Bommasani, Rishi et al. (2021). "On the opportunities and risks of foundation models". In: *arXiv preprint arXiv:2108.07258*.

Boylan, Jack, Chris Hokamp, and Demian Gholipour Ghalandari (2025). "GLiREL-generalist model for zero-shot relation extraction". In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for*

Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 8230–8245.

Brazzo, Laura and Reto Speck (2018). "Holocaust Research and Archives in the Digital Age: Introduction". In: *Quest. Issues in Contemporary Jewish History* 13, pp. V–XIII.

Brown, Tom B. et al. (2020). *Language Models are Few-Shot Learners*. arXiv: 2005.14165 [cs.CL]. URL: https://arxiv.org/abs/2005.14165.

Byrne, William et al. (2004). "Automatic recognition of spontaneous speech for access to multilingual oral history archives". In: *IEEE Transactions on Speech and Audio Processing* 12.4, pp. 420–435.

Chalkidis, Ilias et al. (2020). "LEGAL-BERT: The muppets straight out of law school". In: *arXiv preprint arXiv:2010.02559*.

Chang, Yupeng et al. (2024). "A survey on evaluation of large language models". In: *ACM transactions on intelligent systems and technology* 15.3, pp. 1–45.

Charlton, Thomas L, Lois E Myers, and Rebecca Sharpless (2007). *History of oral history: Foundations and methodology*. Bloomsbury Publishing PLC.

Chen, Bohan and Andrea L Bertozzi (2023). "AutoKG: Efficient automated knowledge graph generation for language models". In: *2023 IEEE International Conference on Big Data (BigData)*. IEEE, pp. 3117–3126.

Chen, Ting et al. (2020). "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PmLR, pp. 1597–1607.

Chen, Xiang et al. (2022). "Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction". In: *Proceedings of the ACM Web conference 2022*, pp. 2778–2788.

Choi, Nayoung et al. (2025). *Trustworthy Answers, Messier Data: Bridging the Gap in Low-Resource Retrieval-Augmented Generation for Domain Expert Systems*. arXiv: 2502.19596 [cs.AI]. URL: https://arxiv.org/abs/2502.19596.

Chung, Hyung Won et al. (2022). *Scaling Instruction-Finetuned Language Models*. arXiv: 2210.11416 [cs.LG]. URL: https://arxiv.org/abs/2210.11416.

Clark, Kevin et al. (2020). "Electra: Pre-training text encoders as discriminators rather than generators". In: *arXiv preprint arXiv:2003.10555*.

Conneau, Alexis, Kartikay Khandelwal, et al. (2019). "Unsupervised cross-lingual representation learning at scale". In: *arXiv preprint arXiv:1911.02116*.

Conneau, Alexis and Guillaume Lample (2019). "Cross-lingual language model pretraining". In: *Advances in neural information processing systems* 32.

Crane, Susan A (2006). "Historical subjectivity: A review essay". In: *The Journal of Modern History* 78.2, pp. 434–456.

Cui, Meiji et al. (2017). "A survey on relation extraction". In: *China Conference on Knowledge Graph and Semantic Computing*. Springer, pp. 50–58.

Cunha, Luís Filipe da Costa and José Carlos Ramalho (2021). "NER in Archival Finding Aids". In.

De Leeuw, Daan et al. (2018). "Digital Methods in Holocaust Studies: The European Holocaust Research Infrastructure". In: *2018 IEEE 14th International Conference on e-Science (e-Science)*. IEEE, pp. 58–66.

De Toni, Francesco et al. (2022). "Entities, dates, and languages: Zero-shot on historical texts with t0". In: *arXiv preprint arXiv:2204.05211*.

Dean, Jeffrey (2022). "A golden decade of deep learning: Computing systems & applications". In: *Daedalus* 151.2, pp. 58–74.

Dermentzi, Maria and Hugo Scheithauer (2024). "Repurposing holocaust-related digital scholarly editions to develop multilingual domain-specific named entity recognition tools". In: *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes)LREC-COLING 2024*, pp. 18–28.

Dettmers, Tim et al. (2023). "Qlora: Efficient finetuning of quantized llms". In: *Advances in neural information processing systems* 36, pp. 10088–10115.

Devlin, Jacob et al. (2019). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of the 2019 conference of the*

North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pp. 4171–4186.

Di Pasquale, Ricardo and Soledad Represa (2024). "Empowering Domain-Specific Language Models with Graph-Oriented Databases: A Paradigm Shift in Performance and Model Maintenance". In: *arXiv preprint arXiv:2410.03867*.

Digital Holocaust Memory Project (2023). *Digitising Material Evidence: Guidelines and Digitisation Report*. Tech. rep. University of Sussex. URL: `https://reframe.sussex.ac.uk/digitalholocaustmemory/%20files/2024/07/Digitising-Material-Evidence-Guidelines-%20Digital-Holocaust-Memory-Project-2.pdf`.

Digital Watch Observatory (June 2024). *UNESCO warns of AI's role in distorting Holocaust history*. Accessed: 2026-05-12. URL: `https://dig.watch/updates/unesco-warns-of-ais-role-in-distorting-holocaust-history`.

Dixit, Kalpit and Yaser Al-Onaizan (2019). "Span-level model for relation extraction". In: *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 5308–5314.

Du, Jinhua et al. (2018). "Multi-level structured self-attentions for distantly supervised relation extraction". In: *arXiv preprint arXiv:1809.00699*.

Dunstan, Jocelyn et al. (2024). "A pseudonymized corpus of occupational health narratives for clinical entity recognition in Spanish". In: *BMC Medical Informatics and Decision Making* 24.1, p. 204.

Dutta, Arka et al. (2024). "Down the toxicity rabbit hole: A framework to bias audit large language models with key emphasis on racism, antisemitism, and misogyny". In: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI*, pp. 3–9.

Eberts, Markus and Adrian Ulges (2020). "Span-based joint entity and relation extraction with transformer pre-training". In: *ECAI 2020*. IOS Press, pp. 2006–2013.

Edge, Darren et al. (2024). "From local to global: A graph rag approach to query-focused summarization". In: *arXiv preprint arXiv:2404.16130*.

Ehrmann, Maud et al. (2023). "Named entity recognition and classification in historical documents: A survey". In: *ACM Computing Surveys* 56.2, pp. 1–47.

Etzioni, Oren et al. (2011). "Open information extraction: The second generation." In: *IJCAI*. Vol. 11, pp. 3–10.

Ezeani, Ignatius, Paul Rayson, Ian Gregory, et al. (2023). "Towards an Extensible Framework for Understanding Spatial Narratives". In: *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Geospatial Humanities*. GeoHumanities '23. Hamburg, Germany: Association for Computing Machinery, pp. 1–10. ISBN: 9798400703492. DOI: 10.1145/3615887.3627761. URL: https://doi.org/10.1145/3615887.3627761.

Ezeani, Ignatius, Paul Rayson, and Ian N Gregory (2023). "Extracting Imprecise Geographical and Temporal References from Journey Narratives." In: *Text2Story@ ECIR*, pp. 113–118.

Ezeani, Ignatius, Paul Rayson, Ian N Gregory, et al. (2024). "The Geography of'Fear','Sadness','Anger'and'Joy': Exploring the Emotional Landscapes in the Holocaust Survivors' Testimonies." In: *Text2Story@ ECIR*, pp. 93–103.

Fairclough, Norman (2013). *Critical discourse analysis: The critical study of language*. Routledge.

Fan, Lizhou and Todd Presner (2022). "Algorithmic Close Reading: Using Semantic Triplets to Index and Analyze Agency in Holocaust Testimonies." In: *DHQ: Digital Humanities Quarterly* 16.3.

Felman, Shoshana and Dori Laub (1992). *Testimony: Crises of witnessing in literature, psychoanalysis, and history*. Taylor & Francis.

Fu, Zihao et al. (2023). *Decoder-Only or Encoder-Decoder? Interpreting Language Model as a Regularized Encoder-Decoder*. arXiv: 2304.04052 [cs.CL]. URL: https://arxiv.org/abs/2304.04052.

Gao, Yunfan et al. (2023). "Retrieval-augmented generation for large language models: A survey". In: *arXiv preprint arXiv:2312.10997* 2.1.

Gee, James Paul (2014). *An introduction to discourse analysis: Theory and method*. routledge.

Gharagozlou, Hamid et al. (2023). "Semantic relation extraction: a review of approaches, datasets, and evaluation methods with looking at the methods and datasets in the Persian language". In: *ACM Transactions on Asian and Low-Resource Language Information Processing* 22.7, pp. 1–29.

González-Gallardo, Carlos-Emiliano et al. (2023). "Yes but.. can chatgpt identify entities in historical documents?" In: *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, pp. 184–189.

Greenspan, Henry et al. (2014). "Engaging survivors: Assessing 'testimony'and 'trauma'as foundational concepts". In: *Dapim: Studies on the Holocaust* 28.3, pp. 190–226.

Gref, Michael et al. (2022). *A Study on the Ambiguity in Human Annotation of German Oral History Interviews for Perceived Emotion Recognition and Sentiment Analysis*. arXiv: 2201.06868 [eess.AS]. URL: https://arxiv.org/abs/2201.06868.

Gritta, Milan, Mohammad Taher Pilehvar, and Nigel Collier (2020). "A pragmatic guide to geoparsing evaluation: Toponyms, Named Entity Recognition and pragmatics". In: *Language resources and evaluation* 54, pp. 683–712.

Gruber, Ivan et al. (2024). "Multi-label Classification and Named Entity Recognition for Historical Documents". In: *International Conference on the Dynamics of Information Systems*. Springer, pp. 24–34.

Gururangan, Suchin et al. (2020). "Don't stop pretraining: Adapt language models to domains and tasks". In: *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 8342–8360.

Halliday, Michael Alexander Kirkwood (1989). "Spoken and written language". In.

Han, Xiaochuang and Jacob Eisenstein (2019). "Unsupervised Domain Adaptation of Contextualized Embeddings: A Case Study in Early Modern English". In: *CoRR* abs/1904.02817. arXiv: 1904.02817. URL: http://arxiv.org/abs/1904.02817.

Han, Xu et al. (2022). "Ptr: Prompt tuning with rules for text classification". In: *AI Open* 3, pp. 182–192.

Haris, Erum, Anthony G Cohn, and John G Stell (2024). "Exploring spatial representations in the historical lake district texts with llm-based relation extraction". In: *arXiv preprint arXiv:2406.14336*.

Hatmi, Mohamed et al. (2013). "Named Entity Recognition in Speech Transcripts following an Extended Taxonomy." In: *SLAM@ INTERSPEECH*, pp. 61–65.

Hilberg, Raul (2003). *The destruction of the European Jews*. Yale University Press.

Hiltmann, Torsten et al. (2025). "NER4all or Context is All You Need: Using LLMs for low-effort, high-performance NER on historical texts. A humanities informed approach". In: *arXiv preprint arXiv:2502.04351*.

Hirsch, Marianne and Leo Spitzer (2009). "The witness in the archive: Holocaust studies/memory studies". In: *Memory Studies* 2.2, pp. 151–170.

Hoffer, Elad and Nir Ailon (2015). "Deep metric learning using triplet network". In: *International Workshop on Similarity-Based Pattern Recognition*. Springer, pp. 84–92.

Hogan, Aidan et al. (July 2021). "Knowledge Graphs". In: *ACM Computing Surveys* 54.4, pp. 1–37. ISSN: 1557-7341. DOI: 10.1145/3447772. URL: http://dx.doi.org/10.1145/3447772.

Honnibal, Matthew and Ines Montani (2017). "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing". To appear.

Hosseini, Kasra et al. (2021). "Neural language models for nineteenth-century english". In: *arXiv preprint arXiv:2105.11321*.

Houlsby, Neil et al. (Sept. 2019). "Parameter-Efficient Transfer Learning for NLP". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by

Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 2790–2799. URL: https://proceedings.mlr.press/v97/houlsby19a.html.

Hsu, Pei-Ling et al. (2015). "Mining various semantic relationships from unstructured user-generated web data". In: *Journal of Web Semantics* 31, pp. 27–38.

Hu, Edward J et al. (2022). "Lora: Low-rank adaptation of large language models." In: *ICLR* 1.2, p. 3.

Hu, Xuke, Jens Kersten, and Friederike Klan (2025). "Scalable Toponym Resolution with LLMs: Accuracy and Speed Optimizations". In: *GeoExT 2025: Third International Workshop on Geographic Information Extraction from Texts at ECIR 2025*. Lucca, Italy. URL: https://ceur-ws.org/Vol-3969/paper6.pdf.

Huang, Lei et al. (2025). "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions". In: *ACM Transactions on Information Systems* 43.2, pp. 1–55.

Hudson, Natalie Florea and Michael J Butler (2010). "The state of experimental research in IR: An analytical survey". In: *International Studies Review* 12.2, pp. 165–192.

Ifergan, Maxim et al. (2024). "Identifying narrative patterns and outliers in holocaust testimonies using topic modeling". In: *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) LREC-COLING 2024*, pp. 44–52.

Ionescu, Arleen and Simona Mitroiu (2023). "Holocaust Narratives in the Post-Testimonial Era: Introduction". In: *Parallax* 29.1, pp. 1–13.

Jadon, Aryan and Avinash Patil (2024). "A comprehensive survey of evaluation techniques for recommendation systems". In: *International Conference on Computation of Artificial Intelligence & Machine Learning*. Springer, pp. 281–304.

Jaff, Daban Q (2025). "CORHOH: Text corpus of holocaust oral histories". In: *Data in Brief* 59, p. 111426.

Ji, Bin et al. (2020). "Span-based joint entity and relation extraction with attention-based span-specific and contextual semantic representations". In: *Proceedings of the 28th international conference on computational linguistics*, pp. 88–99.

Ji, Shaoxiong et al. (2022). "A Survey on Knowledge Graphs: Representation, Acquisition, and Applications". In: *IEEE Transactions on Neural Networks and Learning Systems* 33.2. ISSN: 2162-2388. DOI: 10.1109/tnnls.2021.3070843. URL: http://dx.doi.org/10.1109/TNNLS.2021.3070843.

Ji, Ziwei et al. (2023). "Survey of hallucination in natural language generation". In: *ACM computing surveys* 55.12, pp. 1–38.

Jiang, Albert Q. et al. (2023). *Mistral 7B*. arXiv: 2310.06825 [cs.CL]. URL: https://arxiv.org/abs/2310.06825.

Jiang, Wen (2024). "A Method for Ancient Book Named Entity Recognition Based on BERT-Global Pointer". In: *International Journal of Computer Science and Information Technology*.

Kalai, Adam Tauman and Santosh S Vempala (2024). "Calibrated language models must hallucinate". In: *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pp. 160–171.

Karsvall, Olof and Lars Borin (2018). "SDHK meets NER: Linking Place Names with Medieval Charters and Historical Maps." In: *DHN*, pp. 38–50.

Kekki, Minna-Kerttu M (2024). "Collective memory as sedimentations of collective experience: phenomenological analysis of post-Soviet Europe". In: *Journal of the British Society for Phenomenology* 55.4, pp. 289–307.

Keraghel, Imed, Stanislas Morbieu, and Mohamed Nadif (2024). "A survey on recent advances in named entity recognition". In: *arXiv preprint arXiv:2401.10825*.

Khosla, Prannay et al. (2020). "Supervised contrastive learning". In: *Advances in neural information processing systems* 33, pp. 18661–18673.

Knowles, Anne Kelly, Tim Cole, and Alberto Giordano (2014). *Geographies of the Holocaust*. Indiana University Press.

Konys, Agnieszka and Zygmunt Drążek (2020). "Ontology Learning Approaches to Provide Domain-Specific Knowledge Base". In: *Procedia Computer Science* 176. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES2020, pp. 3324–3334. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2020.09.065. URL: https://www.sciencedirect.com/science/article/pii/S1877050920319608.

Kovács, Éva (2018). "Testimonies in the Digital Age–New Challenges in Research, Academia and Archives". In: *Werner Dreier, Angelika Laumer und Moritz Wein (Hg.): Interactions: Explorations of good practice in educational work with video testimonies of victims of National Socialism, Berlin*, pp. 76–89.

Kraft, Robert N (2004). "Emotional Memory in Survivors of the Holocaust: A Qualitative Study of Oral Testimony." In.

Kumar, Aman and Binil Starly (2022). "FabNER: information extraction from manufacturing process science domain literature using named entity recognition". In: *Journal of Intelligent Manufacturing* 33.8, pp. 2393–2407.

Langer, Lawrence L (1993). *Holocaust testimonies: The ruins of memory*. Yale University Press.

Laub, Dori (2013). "Bearing witness or the vicissitudes of listening". In: *Testimony*. Routledge, pp. 57–74.

Laub, Dori and Nanette Auerhahn (2017). "Knowing and not knowing: Forms of traumatic memory 1". In: *Psychoanalysis and Holocaust Testimony*. Routledge, pp. 32–42.

Lauriola, Ivano, Alberto Lavelli, and Fabio Aiolli (2022). "An introduction to deep learning in natural language processing: Models, techniques, and tools". In: *Neurocomputing* 470, pp. 443–456.

Lee, Benjamin Charles Germain (2019). "Machine learning, template matching, and the International Tracing Service digital archive: Automating the retrieval of

death certificate reference cards from 40 million document scans". In: *Digital Scholarship in the Humanities* 34.3, pp. 513–535.

Lee, Jinhyuk et al. (2020). "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36.4, pp. 1234–1240.

Lee, Seanie et al. (2023). *Self-Distillation for Further Pre-training of Transformers*. arXiv: 2210.02871 [cs.CV]. URL: https://arxiv.org/abs/2210.02871.

Lee, Woong Ki et al. (2012). "Open information extraction for SOV language based on entity-predicate pair detection". In: *Proceedings of COLING 2012: Demonstration Papers*, pp. 305–312.

Leitner, Elena, Georg Rehm, and Julian Moreno-Schneider (2019). "Fine-grained named entity recognition in legal documents". In: *International conference on semantic systems*. Springer, pp. 272–287.

Levy, Omer et al. (2017). "Zero-shot relation extraction via reading comprehension". In: *arXiv preprint arXiv:1706.04115*.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. (2020a). "Retrieval-augmented generation for knowledge-intensive nlp tasks". In: *Advances in neural information processing systems* 33, pp. 9459–9474.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. (2020b). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 9459–9474.

Li, Diyou et al. (2025). "An improved two-stage zero-shot relation triplet extraction model with hybrid cross-entropy loss and discriminative reranking". In: *Expert Systems with Applications* 265, p. 126077. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2024.126077. URL: https://www.sciencedirect.com/science/article/pii/S0957417424029440.

Li, Huaxia et al. (2025). "Extracting financial data from unstructured sources: Leveraging large language models". In: *Journal of Information Systems* 39.1, pp. 135–156.

Li, Minzhi et al. (2023). "Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation". In: *arXiv preprint arXiv:2310.15638*.

Li, Xiaoya et al. (2019). "Entity-relation extraction as multi-turn question answering". In: *arXiv preprint arXiv:1905.05529*.

Lian, Zheng et al. (2025). "Mer 2025: When affective computing meets large language models". In: *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13837–13842.

Liu, Boyu (2002). "Comparative analysis of encoder-only decoder-only and encoder-decoder language models". In: *International Conference on Data Science and Engineering*.

Liu, Haokun et al. (2022). "Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning". In: *Advances in Neural Information Processing Systems* 35, pp. 1950–1965.

Liu, Jiangfeng et al. (2024). *From ChatGPT, DALL-E 3 to Sora: How has Generative AI Changed Digital Humanities Research and Services?* arXiv: 2404.18518 [cs.DL]. URL: https://arxiv.org/abs/2404.18518.

Liu, Ruicheng et al. (2023). "A brief survey on recent advances in coreference resolution". In: *Artificial Intelligence Review* 56.12, pp. 14439–14481.

Liu, Yinhan et al. (2019). "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692*.

Loshchilov, Ilya and Frank Hutter (2017). "Decoupled weight decay regularization". In: *arXiv preprint arXiv:1711.05101*.

Luxburg, Ulrike von (2007). "A Tutorial on Spectral Clustering". In: *CoRR* abs/0711.0189. arXiv: 0711.0189. URL: http://arxiv.org/abs/0711.0189.

Lynch, Renee et al. (2023). "'The tears don't give you funding': data neocolonialism in development in the Global South". In: *Third World Quarterly* 44.5, pp. 911–929.

Makhortykh, Mykola (2024). *AI and the Holocaust: rewriting history? The impact of artificial intelligence on understanding the Holocaust*. https://doi.org/10.54675/ZHJC6844.

Makhortykh, Mykola, Yehor Lyebyedyev, and Daniel Kravtsov (2021). "Past Is another resource: remembering the 70th anniversary of the victory day on livejournal". In: *Nationalities Papers* 49.2, pp. 375–388.

Manjavacas, Enrique and Lauren Fonteyn (2021). "Macberth: Development and evaluation of a historically pre-trained language model for english (1450-1950)". In: *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pp. 23–36.

Mattingly, William (2025). *ushmm dataset*. https://huggingface.co/datasets/wjbmattingly/ushmm-testimonies? Accessed: 2026-3-25.

Mei, Zhiyu et al. (2025). *ReaL: Efficient RLHF Training of Large Language Models with Parameter Reallocation*. arXiv: 2406.14088 [cs.DC]. URL: https://arxiv.org/abs/2406.14088.

Mienye, Ibomoiye Domor et al. (2025). "Large language models: an overview of foundational architectures, recent trends, and a new taxonomy". In: *Discover Applied Sciences* 7.9, p. 1027.

Mitkov, Ruslan (2022). *The Oxford handbook of computational linguistics*. Oxford university press.

Miwa, Makoto and Mohit Bansal (2016). "End-to-end relation extraction using lstms on sequences and tree structures". In: *arXiv preprint arXiv:1601.00770*.

Morrissey, Charles T (2002). "On oral history interviewing". In: *The oral history reader*. Routledge, pp. 121–127.

Naik, Dishita, Ishita Naik, and Nitin Naik (2024). "Decoder-only transformers: the brains behind generative AI, large language models and large multimodal

models". In: *The International Conference on Computing, Communication, Cybersecurity & AI*. Springer, pp. 315–331.

Nayak, Tapas and Hwee Tou Ng (2020). "Effective modeling of encoder-decoder architecture for joint entity and relation extraction". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 05, pp. 8528–8535.

Nenno, Sami (2024). "Bootstrapping public entities. Domain-specific NER for public speakers". In: *Communication Methods and Measures*, pp. 1–26.

Neudecker, Clemens (2016). "An open corpus for named entity recognition in historic newspapers". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 4348–4352.

Nguyen, Minh and Zhou Yu (July 2021). "Improving Named Entity Recognition in Spoken Dialog Systems by Context and Speech Pattern Modeling". In: *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Haizhou Li et al. Singapore and Online: Association for Computational Linguistics, pp. 45–55. DOI: 10.18653/v1/2021.sigdial-1.6. URL: https://aclanthology.org/2021.sigdial-1.6/.

Nightingale, Sophie J, Kimberley A Wade, and Derrick G Watson (2017). "Can people identify original and manipulated photos of real-world scenes?" In: *Cognitive research: principles and implications* 2.1, p. 30.

O'Donoghue, Samuel (2021). "figurations of suffering in concentration camp testimony". In: *comparative literature studies* 58.4, pp. 807–835.

Oord, Aaron van den, Yazhe Li, and Oriol Vinyals (2018). "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748*.

Ouyang, Long et al. (2022). *Training language models to follow instructions with human feedback*. arXiv: 2203.02155 [cs.CL]. URL: https://arxiv.org/abs/2203.02155.

Pai, Liu et al. (2024). "A survey on open information extraction from rule-based model to large language model". In: *Findings of the association for computational linguistics: EMNLP 2024*, pp. 9586–9608.

Pakhale, Kalyani (2023). *Comprehensive Overview of Named Entity Recognition: Models, Domain-Specific Applications and Challenges*. arXiv: 2309.14084 [cs.CL]. URL: https://arxiv.org/abs/2309.14084.

Palmero Aprosio, Alessio, Stefano Menini, and Sara Tonelli (2022). "BERToldo, the Historical BERT for Italian". In: *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*.

Papanikolaou, Yannis, Ian Roberts, and Andrea Pierleoni (2019). "Deep bidirectional transformers for relation extraction without supervision". In: *arXiv preprint arXiv:1911.00313*.

Paszke, Adam et al. (2019). "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32.

Pawar, Sachin, Girish K Palshikar, and Pushpak Bhattacharyya (2017). "Relation extraction: A survey". In: *arXiv preprint arXiv:1712.05191*.

Pearson, Cole, Naeem Seliya, and Rushit Dave (2021). "Named Entity Recognition in Unstructured Medical Text Documents". In: *CoRR* abs/2110.15732. arXiv: 2110.15732. URL: https://arxiv.org/abs/2110.15732.

Pennebaker, James W (2011). "The secret life of pronouns". In: *New Scientist* 211.2828, pp. 42–45.

Perera, Nadeesha, Matthias Dehmer, and Frank Emmert-Streib (2020). "Named entity recognition and relation detection for biomedical information extraction". In: *Frontiers in cell and developmental biology* 8, p. 673.

Pessanha, Francisca and Almila Akdag Salah (Dec. 2021). "A Computational Look at Oral History Archives". In: 15.1. ISSN: 1556-4673. DOI: 10.1145/3477605. URL: https://doi.org/10.1145/3477605.

Pfeiffer, Jonas et al. (Oct. 2020). "AdapterHub: A Framework for Adapting Transformers". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Qun Liu and David Schlangen. Online: Association for Computational Linguistics, pp. 46–54.

DOI: 10.18653/v1/2020.emnlp-demos.7. URL: https://aclanthology.org/2020.emnlp-demos.7/.

Picheny, Michael, Zoltán Tüske, et al. (2019). "Challenging the boundaries of speech recognition: the MALACH corpus". In: *arXiv preprint arXiv:1908.03455*.

Picheny, Michael, Zóltan Tüske, et al. (2019). *Challenging the Boundaries of Speech Recognition: The MALACH Corpus*. arXiv: 1908.03455 [cs.CL]. URL: https://arxiv.org/abs/1908.03455.

Plutchik, Robert (1980). "A general psychoevolutionary theory of emotion". In: *Theories of emotion*. Elsevier, pp. 3–33.

Pornprasit, Chanathip and Chakkrit Tantithamthavorn (2024). "Fine-tuning and prompt engineering for large language models-based code review automation". In: *Information and Software Technology* 175, p. 107523. ISSN: 0950-5849. DOI: https://doi.org/10.1016/j.infsof.2024.107523. URL: https://www.sciencedirect.com/science/article/pii/S0950584924001289.

Poso, Venla et al. (2023). "Untapped data resources. Applying NER for historical archival records of state authorities". In: *Digital Humanities in the Nordic and Baltic Countries Publications* 5.1, pp. 55–69.

Povey, Daniel, Brian Kingsbury, et al. (2005). "fMPE: Discriminatively trained features for speech recognition". In: *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* Vol. 1. IEEE, pp. I–961.

Povey, Daniel and Philip C Woodland (2002). "Minimum phone error and I-smoothing for improved discriminative training". In: *2002 IEEE international conference on acoustics, speech, and signal processing*. Vol. 1. IEEE, pp. I–105.

Presner, Todd (2024). *Ethics of the algorithm: Digital humanities and Holocaust memory*. Princeton University Press.

Psutka, Josef V, Aleš Pražák, and Jan Vaněk (2021). "Recognition of heavily accented and emotional speech of English and Czech Holocaust survivors

using various DNN architectures". In: *International Conference on Speech and Computer*. Springer, pp. 553–564.

Qi, Peng et al. (2020). "Stanza: A Python natural language processing toolkit for many human languages". In: *arXiv preprint arXiv:2003.07082*.

Qiu, Xipeng et al. (2020). "Pre-trained models for natural language processing: A survey". In: *Science China technological sciences* 63.10, pp. 1872–1897.

Qu, Meng et al. (2020). "Few-shot relation extraction via bayesian meta-learning on relation graphs". In: *International conference on machine learning*. PMLR, pp. 7867–7876.

Radford, Alec et al. (2018). *Improving language understanding by generative pre-training.(2018)*.

Raffel, Colin et al. (2020). "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *Journal of machine learning research* 21.140, pp. 1–67.

Ram, Ori et al. (2023). "In-context retrieval-augmented language models". In: *Transactions of the Association for Computational Linguistics* 11, pp. 1316–1331.

Ramshaw, Lance A and Mitchell P Marcus (1999). "Text chunking using transformation-based learning". In: *Natural language processing using very large corpora*, pp. 157–176.

Ranasinghe, Tharindu, **Isuri Anuradha**, et al. (2025). "Sold: Sinhala offensive language dataset". In: *Language Resources and Evaluation* 59.1, pp. 297–337.

Ranasinghe, Tharindu, Hansi Hettiarachchi, et al. (2025). "Sinhala encoder-only language models and evaluation". In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8623–8636.

Ren, Yujie, Niklas Gruhlke, and Anne Lauscher (2025). "Detecting Hallucinations in Authentic LLM-Human Interactions". In: *arXiv preprint arXiv:2510.10539*.

Reynolds, Laria and Kyle McDonell (2021). "Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm". In: New York, NY, USA:

Association for Computing Machinery. ISBN: 9781450380959. DOI: 10.1145/3411763.3451760. URL: https://doi.org/10.1145/3411763.3451760.

Riessman, Catherine Kohler (1993). "Doing narrative analysis". In: *Narrative Analysis. London: Sage Publications*.

Roberts, Jennafer Shae and Laura N Montoya (2022). "Decolonisation, global data law, and Indigenous data sovereignty". In: *arXiv preprint arXiv:2208.04700*.

Robinson, Joshua et al. (2020). "Contrastive learning with hard negative samples". In: *arXiv preprint arXiv:2010.04592*.

Roca Lizarazu, Maria (2017). "Finding the Holocaust in metaphor: renegotiations of trauma in contemporary German-and Austrian-Jewish literature". PhD thesis. University of Warwick.

Rosenthal, Aharona (2025). ""Can AI make Hitler cry?" exploring the use of AI in Holocaust education across four generations". In: *AI and Ethics*, pp. 1–22.

Rousseeuw, Peter J (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: *Journal of computational and applied mathematics* 20, pp. 53–65.

Ruokolainen, Teemu and Kimmo Kettunen (2018). "À la recherche du nom perdu–searching for named entities with Stanford NER in a Finnish historical newspaper and journal collection". In: *13th IAPR International Workshop on Document Analysis Systems*, pp. 1–2.

Sá, Breno Dourado, Ticiana Coelho Da Silva, and José Antônio Fernandes de Macêdo (2022). "Enhancing Geocoding of Adjectival Toponyms with Heuristics". In: *Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences*, pp. 37–45.

Sagi, Tomer et al. (2016). "Multi-source uncertain entity resolution at yad vashem: Transforming holocaust victim reports into people". In: *Proceedings of the 2016 International Conference on Management of Data*, pp. 807–819.

Sahoo, Pranab et al. (2024). "A systematic survey of prompt engineering in large language models: Techniques and applications". In: *arXiv preprint arXiv:2402.07927* 1.

Sanjeev, Suvansh and Anton Troynikov (May 2024). *Embedding Adapters*. Tech. rep. Accessed on July 8, 2024. Chroma. URL: https://research.trychroma.com/embedding-adapters.

Santos, Joaquim et al. (2024). "Named entity recognition specialised for Portuguese 18th-century History research". In: *International Conference on Computational Processing of the Portuguese Language*. URL: https://api.semanticscholar.org/CorpusID:268240767.

Sarker, Shraboni et al. (2024). *Seventeenth-Century Spanish American Notary Records for Fine-Tuning Spanish Large Language Models*. arXiv: 2406.05812 [cs.CL]. URL: https://arxiv.org/abs/2406.05812.

Schick, Timo, Helmut Schmid, and Hinrich Schütze (2020). "Automatically identifying words that can serve as labels for few-shot text classification". In: *arXiv preprint arXiv:2010.13641*.

Schierman, Kelly (2025). "The Geography of Genocide: Using Machine Learning to Locate Undocumented Mass Graves of the Holocaust". In.

Schiffrin, Deborah (1994). *Approaches to discourse*. Vol. 8. Blackwell Oxford.

Schopf, Tim, Dennis N Schneider, and Florian Matthes (2023). "Efficient domain adaptation of sentence embeddings using adapters". In: *arXiv preprint arXiv:2307.03104*.

Schroff, Florian, Dmitry Kalenichenko, and James Philbin (2015). "Facenet: A unified embedding for face recognition and clustering". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823.

Schubert, Lenhart (2020). "Computational Linguistics". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2020. Metaphysics Research Lab, Stanford University.

Schulhoff, Sander et al. (2024). "The prompt report: A systematic survey of prompt engineering techniques". In: *arXiv preprint arXiv:2406.06608*.

Schweter, Stefan et al. (2022). "hmbert: Historical multilingual language models for named entity recognition". In: *arXiv preprint arXiv:2205.15575*.

Sexton, Anna (2025). "Introducing the legacies and trajectories of trauma to the archival field". In: *Archival Science* 25.1, p. 3.

Sharma, Prem and Sangeeta Magar (2024). "Oral Narratives and the Making of History: Streamlining Past and Understanding Challenges". In: *Journal of Dynamics and Control* 8.8, pp. 258–268.

Shen, Yatian and Xuan-Jing Huang (2016). "Attention-based convolutional neural network for semantic relation extraction". In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 2526–2536.

Sherstinsky, Alex (2020). "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network". In: *Physica D: Nonlinear Phenomena* 404, p. 132306.

Shi, Weijia et al. (2023). "Replug: Retrieval-augmented black-box language models". In: *arXiv preprint arXiv:2301.12652*.

Shi, Yue et al. (2012). "Climf: learning to maximize reciprocal rank with collaborative less-is-more filtering". In: *Proceedings of the sixth ACM conference on Recommender systems*, pp. 139–146.

Shin, Jiho et al. (2023). "Prompt engineering or fine tuning: An empirical assessment of large language models in automated software engineering tasks". In: *arXiv preprint arXiv:2310.10508*.

Shopes, Linda (2007). "Legal and ethical issues in oral history". In: *History of oral history: Foundations and methodology*, pp. 125–159.

Siciliani, Lucia et al. (2024). "OIE4PA: open information extraction for the public administration". In: *Journal of Intelligent Information Systems* 62.1, pp. 273–294.

Singh, Sonit (2018). "Natural language processing for information extraction". In: *arXiv preprint arXiv:1807.02383*.

Smith, Victoria et al. (2023). "Identifying and mitigating privacy risks stemming from language models: A survey". In: *arXiv preprint arXiv:2310.01424*.

Soares, Livio Baldini et al. (2019). "Matching the blanks: Distributional similarity for relation learning". In: *arXiv preprint arXiv:1906.03158*.

Son, Junyoung et al. (2022). "GRASP: Guiding model with RelAtional semantics using prompt for dialogue relation extraction". In: *arXiv preprint arXiv:2208.12494*.

Staab, Robin et al. (2024). "Beyond memorization: Violating privacy via inference with large language models". In: *International Conference on Learning Representations*. Vol. 2024, pp. 33832–33878.

Ståhl, Niclas and Lisa Weimann (2022). "Identifying wetland areas in historical maps using deep convolutional neural networks". In: *Ecological Informatics* 68, p. 101557. ISSN: 1574-9541. DOI: https://doi.org/10.1016/j.ecoinf.2022.101557. URL: https://www.sciencedirect.com/science/article/pii/S1574954122000061.

Stone, Dan (2010). *Histories of the Holocaust*. Oxford University Press.

Subramaniam, L. Venkata et al. (2009). "A survey of types of text noise and techniques to handle noisy text". In: *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*. AND '09. Barcelona, Spain: Association for Computing Machinery, pp. 115–122. ISBN: 9781605584966. DOI: 10.1145/1568296.1568315. URL: https://doi.org/10.1145/1568296.1568315.

Tannen, Deborah (1982). "Spoken and written language: Exploring orality and literacy". In.

The Times of Israel Staff (May 2025). *Auschwitz museum sounds alarm over "harmful" AI images of Holocaust victims*. Accessed: 2025-11-27. URL: https://www.timesofisrael.com/auschwitz-museum-sounds-alarm-over-harmful-ai-images-of-holocaust-victims/ (visited on 11/27/2025).

Timmons, Adela C et al. (2023). "A call to action on assessing and mitigating bias in artificial intelligence applications for mental health". In: *Perspectives on Psychological Science* 18.5, pp. 1062–1096.

Tiribelli, Simona et al. (2024). "Ethics of artificial intelligence for cultural heritage: Opportunities and challenges". In: *IEEE Transactions on Technology and Society* 5.3, pp. 293–305.

Toni, Francesco De et al. (2022). *Entities, Dates, and Languages: Zero-Shot on Historical Texts with T0*. arXiv: 2204.05211 [cs.CL]. URL: https://arxiv.org/abs/2204.05211.

Torfi, Amirsina et al. (2021). *Natural Language Processing Advancements By Deep Learning: A Survey*. arXiv: 2003.01200 [cs.CL]. URL: https://arxiv.org/abs/2003.01200.

Touvron, Hugo et al. (2023). "Llama: Open and efficient foundation language models". In: *arXiv preprint arXiv:2302.13971*.

United States Holocaust Memorial Museum (2020). *Introduction to the Holocaust*. USHMM. URL: https://encyclopedia.ushmm.org/content/en/article/introduction-to-the-holocaust.

Vaswani, Ashish et al. (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30.

Venkit, Pranav Narayanan et al. (2024). "Search Engines in an AI Era: The False Promise of Factual and Verifiable Source-Cited Responses". In: *arXiv preprint arXiv:2410.22349*.

Villena, Fabián, Felipe Bravo-Marquez, and Jocelyn Dunstan (2025). "NLP modeling recommendations for restricted data availability in clinical settings". In: *BMC Medical Informatics and Decision Making* 25.1, p. 116.

Wagner, Eitan, Renana Keydar, and Omri Abend (Dec. 2023). "Event-Location Tracking in Narratives: A Case Study on Holocaust Testimonies". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for

Computational Linguistics, pp. 8789–8805. DOI: 10.18653/v1/2023.emnlp-main.544. URL: https://aclanthology.org/2023.emnlp-main.544/.

Wang, Jiahui, Kun Yue, and Liang Duan (2023). "Models and techniques for domain relation extraction: a survey". In: *Journal of Data Science and Intelligent Systems* 1.2, pp. 65–82.

Wang, Liang et al. (2023). "Improving text embeddings with large language models". In: *arXiv preprint arXiv:2401.00368*.

Wang, Xiaoxiong and Jianpeng Hu (2023). "An Open Relation Extraction Method for Domain Text Based on Hybrid Supervised Learning". In: *Applied Sciences* 13.5. ISSN: 2076-3417. DOI: 10.3390/app13052962. URL: https://www.mdpi.com/2076-3417/13/5/2962.

Warner, Benjamin et al. (2025). "Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference". In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2526–2547.

Weerasinghe, Ruvan, **Isuri Anuradha**, and Deshan Sumanathilaka (2025). "Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages". In: *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*.

Wei, Jason, Maarten Bosma, et al. (2022). *Finetuned Language Models Are Zero-Shot Learners*. arXiv: 2109.01652 [cs.CL]. URL: https://arxiv.org/abs/2109.01652.

Wei, Jason, Xuezhi Wang, et al. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. arXiv: 2201.11903 [cs.CL]. URL: https://arxiv.org/abs/2201.11903.

Wei, Qiang et al. (2020). "Relation extraction from clinical narratives using pre-trained language models". In: *AMIA annual symposium proceedings*. Vol. 2019, p. 1236.

Williams, Lowri et al. (Apr. 2024). "Topic Modelling: Going beyond Token Outputs". In: *Big Data and Cognitive Computing* 8.5, p. 44. ISSN: 2504-2289. DOI: 10.3390/bdcc8050044. URL: http://dx.doi.org/10.3390/bdcc8050044.

Wilson, Caitlin (2023). *Working with NLP and holocaust testimonies*. en. https://clarin.web.ox.ac.uk/article/working-nlp-and-holocaust-testimonies. Accessed: 2026-3-25.

Wittgen, Arne Steffen, Faegheh Hasibi, and Serge Thill (2023). "Relation Extraction using Few-Shot Entailment on Conversational Data". In.

Wu, Kehan et al. (2023). "Deep learning models for spatial relation extraction in text". In: *Geo-spatial Information Science* 26.1, pp. 58–70.

Wu, Yonghui et al. (2018). "Clinical named entity recognition using deep learning models". In: *AMIA annual symposium proceedings*. Vol. 2017, p. 1812.

Wynne, Martin (May 2023). *Using Holocaust testimonies as Research Data*. URL: https://www.clarin.ac.uk/article/using-holocaust-testimonies-research-data.

Xie, Xin et al. (2022). "Lambdakg: A library for pre-trained language model-based knowledge graph embeddings". In: *arXiv preprint arXiv:2210.00305*.

Xu, Kun et al. (2015). "Semantic relation classification via convolutional neural networks with simple negative sampling". In: *arXiv preprint arXiv:1506.07650*.

Xu, Weiwen et al. (2022). "ConReader: Exploring implicit relations in contracts for contract clause extraction". In: *arXiv preprint arXiv:2210.08697*.

Xu, Xin et al. (2022). "Towards realistic low-resource relation extraction: A benchmark with empirical baseline study". In: *arXiv preprint arXiv:2210.10678*.

Yadav, Vikas and Steven Bethard (Aug. 2018). "A Survey on Recent Advances in Named Entity Recognition from Deep Learning models". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Ed. by Emily M. Bender, Leon Derczynski, and Pierre Isabelle. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 2145–2158. URL: https://aclanthology.org/C18-1182/.

Yang, Soyoung et al. (July 2023). "HistRED: A Historical Document-Level Relation Extraction Dataset". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 3207–3224. DOI: 10.18653/v1/2023.acl-long.180. URL: https://aclanthology.org/2023.acl-long.180/.

Yang, Zhilin et al. (2019). "Xlnet: Generalized autoregressive pretraining for language understanding". In: *Advances in neural information processing systems* 32.

Ye, Wei et al. (2019). "Exploiting entity BIO tag embeddings and multi-task learning for relation extraction with imbalanced data". In: *arXiv preprint arXiv:1906.08931*.

Zaghloul, Waleed and Silvana Trimi (2017). "Developing an innovative entity extraction method for unstructured data". In: *International Journal of Quality Innovation* 3, pp. 1–10.

Zaratiana, Urchade, Gil Pasternak, et al. (2025). "GLiNER2: Schema-driven multi-task learning for structured information extraction". In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 130–140.

Zaratiana, Urchade, Nadi Tomeh, et al. (2024). "GLiNER: Generalist model for named entity recognition using bidirectional transformer". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5364–5376.

Zelenko, Dmitry, Chinatsu Aone, and Anthony Richardella (2003). "Kernel methods for relation extraction". In: *Journal of machine learning research* 3.Feb, pp. 1083–1106.

Zeng, Daojian et al. (2014). "Relation classification via convolutional deep neural network". In: *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, pp. 2335–2344.

Zhang, Ranran Haoran et al. (2020). "Minimize exposure bias of seq2seq models in joint entity and relation extraction". In: *arXiv preprint arXiv:2009.07503*.

Zhang, Yiqun et al. (2026). "Affective computing in the era of large language models: A survey from the nlp perspective". In: *Knowledge-Based Systems*, p. 115411.

Zhang, Yuzhe and Hong Zhang (2023). "FinBERT–MRC: financial named entity recognition using BERT under the machine reading comprehension paradigm". In: *Neural Processing Letters* 55.6, pp. 7393–7413.

Zhang, Zeyu, Egoitz Laparra, and Steven Bethard (2024). "Improving toponym resolution by predicting attributes to constrain geographical ontology entries". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 35–44.

Zhao, Tianyang et al. (2021). "Asking effective and diverse questions: A machine reading comprehension based framework for joint entity-relation extraction". In: *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pp. 3948–3954.

Zhao, Wayne Xin et al. (2023). "A survey of large language models". In: *arXiv preprint arXiv:2303.18223* 1.2, pp. 1–124.

Zhao, Xiaoyan et al. (2024). "A comprehensive survey on relation extraction: Recent advances and new frontiers". In: *ACM Computing Surveys* 56.11, pp. 1–39.

Zheng, Suncong et al. (2017). "Joint extraction of entities and relations based on a novel tagging scheme". In: *arXiv preprint arXiv:1706.05075*.

Zheng, Ziqiang et al. (2020). "TCMKG: A deep learning based traditional Chinese medicine knowledge graph platform". In: *2020 IEEE international conference on knowledge graph (ICKG)*. IEEE, pp. 560–564.

Zhou, Ce et al. (2023). *A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT*. arXiv: 2302.09419 [cs.AI]. URL: https://arxiv.org/abs/2302.09419.

Zhou, Peng et al. (2017). "Joint extraction of multiple relations and entities by using a hybrid neural network". In: *China National Conference on Chinese Computational Linguistics*. Springer, pp. 135–146.

Zhou, Shaowen et al. (2022). "A Survey on Neural Open Information Extraction: Current Status and Future Directions". In: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 5694–5701. DOI: 10.24963/ijcai.2022/793.

Zhu, Yuqi et al. (2024). *LLMs for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities*. arXiv: 2305.13168 [cs.CL]. URL: https://arxiv.org/abs/2305.13168.

Zhuang, Liu et al. (2021). "A robustly optimized BERT pre-training approach with post-training". In: *Proceedings of the 20th chinese national conference on computational linguistics*, pp. 1218–1227.