When Clinical AI Hallucinates or Appears To: A Reflective Account of Human-AI Diagnostic Interaction in General Practice

Yuhao Sun
Lancaster University
Lancaster, United Kingdom
University of Edinburgh
Edinburgh, United Kingdom
yuhao.sun@lancaster.ac.uk

Junsheng Sun
Department of General Practice
Longgang Central Hospital of Shenzhen
Shenzhen, China
Shenzhen Clinical College of Medicine
Guangzhou University of Chinese Medicine
Shenzhen, China
sjunsheng@gzucm.edu.cn

Abstract

Large language models (LLMs) are increasingly used as informal decision support in clinical practice, yet strong benchmark performance does not directly translate to real-world work where information is incomplete, evolving, and distributed across heterogeneous records. We present this co-authored reflective case study of in-the-wild LLM use by a senior general practitioner across outpatient and inpatient settings. Analysing three clinical vignettes, we identify how hallucination-like breakdowns can arise from both factual errors and opaque evidence blending: the model synthesises claims across various record types without making provenance visible, leading grounded details to appear fabricated and speculative inferences to resemble chart facts. We show how disagreement triggers verification work, shifting cognitive load from clinical reasoning to auditing sources. We conclude with design implications for clinical LLM interfaces, including typed provenance links, separation of retrieved evidence from inference, dynamic case reconstitution, and workflows for productive disagreement.

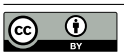
CCS Concepts

• **Human-centered computing → Human computer interaction (HCI)**; **Interaction design process and methods**; • **Applied computing → Health care information systems**.

Keywords

Human-AI Interaction, Clinical Decision-Making, Large Language Models, Hallucination, Evidence Provenance, General Practitioner, Healthcare

ACM Reference Format:
Yuhao Sun and Junsheng Sun. 2026. When Clinical AI Hallucinates or Appears To: A Reflective Account of Human-AI Diagnostic Interaction in General Practice. In *Interactive Health Conference (IH '26), July 05–08, 2026, Porto, Portugal.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3786579.3804942



IH '26, Porto, Portugal

ACM ISBN 979-8-4007-2422-0/26/07
https://doi.org/10.1145/3786579.3804942

1 Introduction

Large language models (LLMs) have demonstrated strong performance on medical question answering and reasoning benchmarks, motivating interest in clinical applications [9, 19]. However, performance in controlled settings does not directly translate to the realities of clinical work, where information is incomplete, evolving, and distributed across heterogeneous sources (patient history, labs, imaging, consultation notes, guidelines), and where accountability and safety requirements are high [1, 2, 10, 12, 18, 22]. Recent studies suggest LLM access can influence clinicians' reasoning processes and outcomes, but results are mixed and highly dependent on task framing and context [7, 11, 12]. For example, a randomised clinical trial in *JAMA Network Open* found that LLM access did not uniformly improve diagnostic reasoning compared with conventional resources, while also highlighting process-level considerations beyond simple accuracy metrics [11].

In parallel, HCI research has emphasised that trustworthy human-AI interaction requires careful design: users need to understand what the system can do, when it may fail, and how to appropriately calibrate reliance [4]. Amershi et al.'s "Guidelines for Human-AI Interaction" provide a foundational framing for designing AI-infused products, including transparency, controllability, and supporting efficient correction [4]. More recent HCI work on precision medicine has highlighted how model-derived, probabilistic outputs in medical contexts challenge conventional assumptions of interpretability, actionability, and user responsibility [20, 21].

Yet clinical use introduces additional complexities: model outputs may be taken as quasi-authoritative; hallucinations can be harmful; and evidence provenance is essential for clinical accountability [12, 15, 19]. In LLM research, hallucination typically denotes outputs that appear plausible and confident but are not supported by the underlying evidence (e.g., fabricated details, incorrect attributions, or unverifiable citations) [14]. In clinical settings, this is particularly risky because such claims can be misread as chart facts, shaping diagnosis, treatment decisions, and documentation [12]. Importantly, hallucination may also be experienced interactionally when evidence is present somewhere in the record, but its provenance is not made visible, making grounded statements appear invented.

While clinical LLMs are increasingly studied through benchmarks and controlled evaluations, clinician-centred accounts of in-the-wild use remain limited – especially those documenting how these systems are appropriated in *real* clinical workflows, including

breakdowns, verification practices, and accountability pressures. We therefore adopt a co-authored reflective case study to surface interactional phenomena that may be difficult to capture in laboratory settings. Rather than treating hallucination purely as a model defect, we examine a broader socio-technical phenomenon: clinicians may experience **perceived hallucination** when LLMs provide confident claims whose source and evidential status are not visible. In dynamic settings, clinicians continuously update their interpretation as new results arrive; LLM responses, by contrast, are often anchored to an incomplete documentation snapshot. This mismatch can generate perceived unreliability even when parts of the output are grounded in the record. This paper makes three main contributions:

(1) An in-the-wild reflective case study of outpatient and inpatient LLM use by a senior clinician working in general practice;
(2) An interactional analysis of hallucination-like breakdowns, focusing on opaque evidence blending across record types;
(3) A set of design implications for clinical LLM interfaces, including provenance, dynamic reconstitution, and productive disagreement.

2 Related Work

We briefly review related work in three areas that motivate our study: (1) clinical LLM use beyond benchmarks, including in-the-wild workflow evidence; (2) LLMs for clinical documentation and multi-document record synthesis, where verifiability and provenance are central; and (3) HCI research on trustworthy human-AI interaction.

LLMs and clinical knowledge. LLMs have been evaluated for clinical knowledge using curated benchmarks and human assessment frameworks. Singhal et al. assessed large models on MultiMedQA and proposed evaluation axes including factuality and potential harm, highlighting that clinical deployment demands more than correct answers [19]. Eriksen et al.'s study examined GPT-4 on complex clinical cases, comparing performance with physicians and illustrating both promise and limitations of case-based reasoning [9].

Decision support and workflow integration. Clinical decision support (CDS) has a long history, and adoption challenges are often socio-technical: usability, integration, and workflow fit matter as much as algorithmic capability. Middleton et al.'s 25-year retrospective highlights that CDS impact depends on implementation and the clinical context of use [17]. These lessons are directly relevant to LLM assistants embedded in office software or electronic health record tools [6].

Transparency, trust, and interaction design. In the field of HCI, trust is not a static user attitude but a situated, evolving judgment shaped by system behaviour and interaction affordances [16]. Amershi et al.'s guidelines for human-AI interaction emphasise transparency about system confidence and supporting user correction loops [4]. In clinical settings, the need for traceability is heightened: clinicians must be able to justify decisions using evidence. When LLMs provide fluent answers without clear provenance, clinicians may interpret outputs as hallucinated – even when partially grounded in records [3, 12].

3 Methods

We adopted a reflective case study approach grounded in clinical reflective practice and interpretive HCI traditions. The second author, a senior general practitioner (GP) and head of department with more than thirty years of clinical experience, documented routine and critical LLM interactions in his clinical workflow. Prior to this role, the GP trained and practised as a respiratory physician. The GP worked across both outpatient and inpatient settings and used the LLM tool as part of everyday clinical reasoning. We then used purposive sampling to select three vignettes that (1) were clinically consequential, (2) represented distinct interaction outcomes (e.g., helpful recall vs. contradiction vs. perceived hallucination due to evidence blending), and (3) provided sufficient traceable context for reflection. We additionally reviewed the remaining documented interactions as a lightweight negative-case analysis; none introduced qualitatively new failure modes beyond the three reported, though we note that this corpus is not exhaustive.

Our dataset comprised four sources: (1) clinician field notes; (2) excerpts/transcripts/screenshots of LLM queries and responses; (3) reconstructed timelines of what information was available at each interaction; and (4) co-analysis discussions and peer debriefing between authors to challenge interpretations and surface alternative explanations. All patient information reported in this paper was de-identified and presented as clinical vignettes. The LLM interactions described here did not involve patient-identifiable information. The study has been approved by the Medical Ethics Committee of Longgang Central Hospital of Shenzhen (reference: 2026ECPJ010), and was conducted in accordance with institutional policies for reflective practice.

The LLM examined in this study was an "AI Assistant" (hereafter, the assistant) embedded within the hospital's office software environment and used in routine clinical work. Although direct access to the external internet was not permitted within the hospital's internal office system, this system included an integrated assistant that, according to the GP, was based on DeepSeek. However, the technical configuration of this system (e.g., whether it involved local deployment or managed external connectivity) was not available to the authors and is therefore not specified further here.

Interactions were documented over a period of three months, during routine outpatient and inpatient clinical work. The assistant was consulted in situations such as diagnostic uncertainty, medication safety checks, or the synthesis of complex patient records. Then, we conducted a structured reflective analysis of each interaction following a 'breakdown-repair' lens from human-AI interaction. For each case, we (1) reconstructed the clinical information environment (what documents were available and when), (2) traced the model's claims to possible sources (e.g., imaging report vs. consult note), (3) identified the user's inference and verification steps, and (4) coded breakdowns by their interactional cause (e.g., contradiction, missing provenance, cross-document evidence blending). We iterated these codes across vignettes to derive recurring mechanisms and design implications.

Table 1: Corpus overview and mapping between observed interaction outcomes and selected vignettes.

Interaction outcome	Observed frequency	Vignette
Helpful recall / augmentation	Common	A
Internal contradiction / safety risk	Occasional	B
Perceived hallucination via provenance opacity / evidence blending	Occasional (high impact)	C

4 Findings

To support readers without a clinical background, we briefly contextualise each vignette in terms of the clinical task and its stakes. We use the vignettes not to evaluate medical correctness, but to illustrate interactional dynamics in everyday clinical decision-making: rapid medication safety checks, managing uncertainty for special populations, and high-stakes diagnostic synthesis across distributed records. Table 1 summarises the broader corpus and the rationale for selecting three analytically rich vignettes.

4.1 Outpatient use: rapid adverse effect reasoning as cognitive offloading

4.1.1 Vignette A (moxifloxacin and fatigue). A middle-aged patient reported persistent fatigue after taking moxifloxacin for a respiratory infection. The assistant was consulted during the consultation to support rapid assessment of whether the symptom could be drug-related. The GP queried[1] whether fatigue is a recognised adverse effect and what safety checks to consider. The assistant returned an answer aligning with known side-effect profiles (fatigue, dizziness, gastrointestinal effects) and suggested basic monitoring and escalation criteria.

4.1.2 Interactional observation. In outpatient contexts, the LLM was most valuable as a 'rapid recall' tool – an external memory that reduces search time. The GP treated outputs as provisional and cross-checked when needed, but the LLM's role came from speed and structure (e.g., listing differential adverse effects and practical next steps). This mode is consistent with a CDS function that supports clinicians' time-constrained decision making rather than replacing judgment [17].

4.2 Special populations: when LLM outputs conflict internally

4.2.1 Vignette B (glucose-6-phosphate dehydrogenase (G6PD) deficiency and medication options). A teenager with G6PD deficiency[2] required symptomatic management for upper respiratory complaints. The assistant was consulted during the consultation as part of the reasoning about medication safety. The assistant initially produced a cautious recommendation, then, in the same response (or a subsequent turn), presented inconsistent statements about whether a specific drug was "not recommended" versus "generally safe," while offering references of uneven quality (mostly publicly available webpages).

4.2.2 Interactional observation. This case illustrates a common LLM risk: superficially authoritative text can contain internal contradictions [15]. The GP responded by treating the assistant as a starting point, then verifying against trusted drug references. For special populations, the GP reported that the lack of access to authoritative databases within the assistant limited its clinical reliability, despite its helpful structuring of considerations (e.g., "watch for haemolysis signs").

4.2.3 Implication. If LLMs are used for medication safety in special populations, interfaces should (a) foreground uncertainty[3]; (b) separate what is 'retrieved evidence' from what is 'model inference'; and (c) make conflicts visible rather than hiding them in a fluent narrative.

4.3 Inpatient multimorbidity: diagnostic disagreement and apparent hallucination via opaque evidence blending

4.3.1 Vignette C (acute transverse myelitis vs. antiphospholipid syndrome). In an inpatient case with complex presentation and evolving results[4], the assistant was used iteratively during the care process, including after new clinical data became available. The assistant initially ranked antiphospholipid syndrome as the top diagnosis and acute transverse myelitis as secondary. After multidisciplinary team (MDT) discussion and external specialist consultation, acute transverse myelitis was supported and treated with corticosteroids.

A critical moment occurred when the assistant, after later reuploads of more records of the patient, provided a confident justification, including a specific magnetic resonance imaging phrase (i.e., "T2 hyperintensity at T12-L1") that the GP did not recall seeing in radiology reports. This was initially perceived as a hallucination: a plausible-sounding imaging finding seemingly invented to support the diagnosis. On investigation, the GP located similar wording in a neurology consultation note. The phrase existed in the record, but its provenance was not clear to the GP at the time of interaction because (1) the record was distributed across documents, and (2) the assistant did not indicate which document type or which date the phrase came from.

[1] Such queries are common in time-pressured outpatient practice, where clinicians must quickly assess whether new symptoms may be drug-related and decide whether monitoring, switching therapy, or escalation is warranted.

[2] G6PD deficiency is a common inherited enzyme disorder in which reduced G6PD activity makes red blood cells more vulnerable to oxidative stress, sometimes leading to acute haemolysis triggered by infections, certain medications, or fava bean consumption [13]. G6PD deficiency constrains medication choices because some drugs can precipitate haemolysis; clinicians therefore rely on authoritative references and careful risk communication, particularly when evidence is ambiguous or conflicting.

[3] In certain clinical scenarios, clinicians may need to choose the least harmful option among imperfect alternatives and engage in shared decision-making with patients. Providing a tentative, evidence-based recommendation can support more informed clinical judgment, even when uncertainty or potential harm exists.

[4] In inpatient care, diagnoses are often revised as new tests arrive, and key evidence is distributed across radiology reports, consultation notes, and handover narratives, making provenance and document type crucial for appropriate evidence weighting.

4.3.2 Interactional observation. This incident is not well captured by the binary 'hallucination vs. non-hallucination' framing, but instead demonstrates a mismatch between evidential presence and evidential legibility. Even when the content is present somewhere in documentation, opaque synthesis can make the output appear fabricated. Conversely, such synthesis can also elevate low-quality or interpretive notes to the same status as formal radiology reports, potentially distorting evidence weighting. The core problem is evidence provenance opacity: the GP cannot see whether a claim is grounded in (a) radiology reports, (b) specialist notes, (c) clinician narrative, or (d) model inference.

4.3.3 Clinician response pattern. The GP did not accept the assistant as decisive; disagreement triggered verification: re-checking reports, searching notes, and consulting external specialists. However, this verification is time-consuming. The cognitive load shifts from clinical reasoning to auditing LLM outputs and record sources – an unintended burden.

5 Discussion

Evidence provenance is a critical interaction requirement in clinicians' engagement with LLM outputs. As illustrated in Figure 1, our findings suggest that the key issue lies in whether their evidential basis remains legible through interaction. Prior HCI work emphasises that AI systems should make it easy for users to understand and appropriately rely on outputs [4]. It also echoes recent research on human-centred AI in healthcare, which emphasises the importance of transparency, workflow integration, and the situated nature of clinical AI use [5, 8]. In clinical contexts, this requires provenance-aware interaction: each key claim should be traceable to a source, and sources should be typed (radiology report vs. consultation note vs. patient-reported history vs. model inference). Without this, clinicians may interpret grounded statements as hallucinations, or more dangerously, treat speculative inferences as documented facts.

Dynamic clinical information challenges static conversational snapshots. Clinical diagnosis unfolds over time; LLM interactions are often episodic snapshots. The same patient, queried at different moments or in new chat windows, can yield different rankings, which clinicians may interpret as inconsistency or unreliability. This aligns with concerns raised by clinical CDS research: tool effectiveness depends on integration with clinical workflows and timely data availability [17].

Productive disagreement should be designed, not improvised. In our cases, clinician-AI disagreement was a productive trigger for verification and escalation (MDT and specialist consultation). However, today this is improvised: clinicians must manually challenge outputs and reconstruct evidence. Clinical LLM interfaces should explicitly support disagreement: making it easy to ask why, see supporting and contradicting evidence, and record clinician judgments for the ongoing case narrative. This would align with HCI principles of controllability and supporting efficient correction [4].

Taken together, these findings point to five design implications for clinical LLMs:

- **Evidence traceability with source typing.** Every key claim should link to a specific source snippet and metadata (e.g., document type, author role, timestamp).
- **Separation of record-grounded facts vs. model inference.** Interface cues should distinguish 'documented' from 'inferred,' reducing the risk of fabricated or speculative details being treated as chart facts.
- **Dynamic case reconstitution.** Provide tooling to periodically rebuild a 'current case state' from evolving data, rather than relying on fragmented conversational history.
- **Disagreement workflows.** Enable structured contestation (e.g., clinician flags "I disagree"), prompting the system to surface counterevidence and uncertainty, and to log the resolution pathway.
- **Quality-aware weighting of evidence.** Radiology reports, lab results, and specialist notes should not be treated as equivalent; interfaces should support weighting and highlight conflicting interpretations.

6 Limitations and Conclusion

Our work is a reflective case study centred on an experienced clinician and a single LLM assistant embedded in local workplace software. As such, the goal is to surface a class of interactional breakdowns that are difficult to observe in benchmark evaluations or decontextualised usability studies, rather than population-level generalisation. The vignettes are shaped by the conventions of one organisational record system (e.g., how documents are fragmented, updated, and re-uploaded) and by an established set of verification, escalation, and documentation routines. These constraints delimit what can be claimed: the study does not estimate the prevalence of failure modes, nor does it compare models or interfaces.

At the same time, the account surfaces two cross-cutting tensions relevant to clinical LLM deployments. First, conversational outputs privilege coherent narrative, while clinical practice depends on graded epistemic status – distinguishing what is observed, reported, inferred, and merely suspected. Second, clinical reasoning is temporally distributed, yet LLM interactions remain episodic. Together, these tensions help explain why grounded content may still be experienced as hallucinated when evidential status and provenance are not legible at the point of interaction.

Future work should therefore treat provenance and temporality as first-class design and evaluation dimensions. This includes (i) interface prototyping that makes evidential status and record provenance legible in situ; (ii) studies across sites and experience levels to examine variation in verification practices, disagreement, and accountability pressures; and (iii) evaluation protocols that go beyond answer accuracy to measure interactional outcomes, such as time-to-verification, error pathways, trust calibration, and documentation quality.

This study reframes hallucination from a purely model-centric defect to an interactional phenomenon shaped by evidential legibility in distributed clinical records. By centring provenance, epistemic status, and temporality as interaction requirements, we hope future HCI research can help steer clinical LLMs towards safer and more accountable use in routine care.

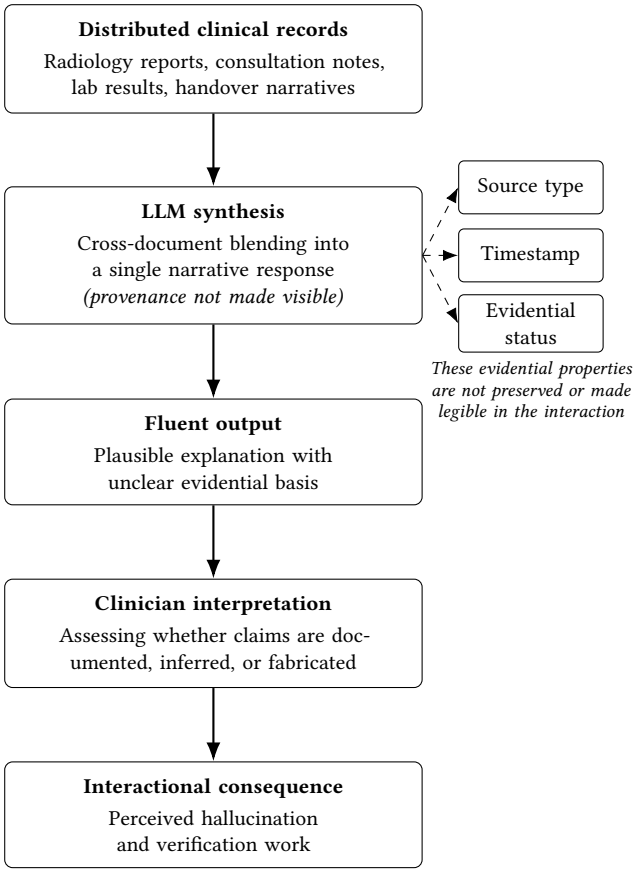


Figure 1: A conceptual account of perceived hallucination in clinical LLM interaction. Information distributed across heterogeneous clinical records is synthesised into a coherent response without preserving visible provenance or evidential status. As a result, grounded content may appear fabricated, while inferred content may be mistaken for chart fact, leading to verification work.

Acknowledgments

This work was supported by the Office of the President, Guangzhou University of Chinese Medicine (2024-278). We acknowledge that this work is grounded in everyday clinical practice and the broader context of patient care, and thank the anonymous reviewers for their constructive feedback.

References

[1] EU Artificial Intelligence Act. 2024. The eu artificial intelligence act. *European Union* (2024).

[2] Monica Agrawal, Irene Y Chen, Freya Gulamali, and Shalmali Joshi. 2025. The evaluation illusion of large language models in medicine. *npj Digital Medicine* 8, 1 (2025), 600.

[3] Mohammad Alkhalaf, Ping Yu, Mengyang Yin, and Chao Deng. 2024. Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records. *Journal of biomedical informatics* 156 (2024), 104662.

[4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[5] Tariq Osman Andersen, Francisco Nunes, Lauren Wilcox, Enrico Coiera, and Yvonne Rogers. 2023. Introduction to the special issue on human-centred AI in healthcare: Challenges appearing in the wild. 12 pages.

[6] Yaara Artsi, Vera Sorin, Benjamin S Glicksberg, Panagiotis Korfiatis, Girish N Nadkarni, and Eyal Klang. 2025. Large language models in real-world clinical workflows: a systematic review of applications and implementation. *Frontiers in Digital Health* 7 (2025), 1659134.

[7] Stephanie Cabral, Daniel Restrepo, Zahir Kanjee, Philip Wilson, Byron Crowe, Raja-Elie Abdulnour, and Adam Rodman. 2024. Clinical reasoning of a generative artificial intelligence model compared with physicians. *JAMA internal medicine* 184, 5 (2024), 581–583.

[8] Lorenzo Corti, Rembrandt Oltmans, Jiwon Jung, Agathe Balayn, Marlies Wijsenbeek, and Jie Yang. 2024. "It Is a Moving Process": Understanding the Evolution of Explainability Needs of Clinicians in Pulmonary Medicine. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–21.

[9] Alexander V Eriksen, Sören Möller, and Jesper Ryg. 2024. Use of GPT-4 to diagnose complex clinical cases. *NEJM AI* 1, 1 (2024).

[10] Geraldine Fitzpatrick and Gunnar Ellingsen. 2013. A review of 25 years of CSCW research in healthcare: contributions, challenges and future agendas. *Computer Supported Cooperative Work (CSCW)* 22, 4 (2013), 609–665.

[11] Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A Cool, Zahir Kanjee, Andrew S Parsons, Neera Ahuja, et al. 2024. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA network open* 7, 10 (2024), e2440969–e2440969.

[12] Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine* 30, 9 (2024), 2613–2622.

[13] Susan J Harcke, Denise Rizzolo, and H Theodore Harcke. 2019. G6PD deficiency: An update. *Jaapa* 32, 11 (2019), 21–26.

[14] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* 43, 2 (2025), 1–55.

[15] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys* 55, 12 (2023), 1–38.

[16] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[17] Blackford Middleton, Dean F Sittig, and Adam Wright. 2016. Clinical decision support: a 25 year retrospective and a 25 year vision. *Yearbook of medical informatics* 25, S 01 (2016), S103–S116.

[18] World Health Organization. 2024. *Ethics and governance of artificial intelligence for health: large multi-modal models. WHO guidance.* World Health Organization.

[19] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (2023), 172–180.

[20] Yuhao Sun, Albert Tenesa, and John Vines. 2025. Human-Precision Medicine Interaction: Public Perceptions of Polygenic Risk Score for Genetic Health Prediction. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–26.

[21] Yuhao Sun, Albert Tenesa, and John Vines. 2026. How Can We Make Precision Medicine (PM) Contestable? Provotyping for PM Service Ecosystems. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–9.

[22] Elham Tabassi. 2023. Artificial intelligence risk management framework (AI RMF 1.0). (2023).