Evaluating chatbot authenticity in simulations of spoken interaction: Demonstrating the utility of corpus-based methods for development and validation

Abstract

This study presents a methodological framework for applying corpus linguistics to systematically evaluate the authenticity of chatbot production in relation to (spoken) production in a general target language use domain. We demonstrate the approach through data drawn from the development cycle of a low-stakes formative assessment system in which learners interact with a ChatGPT-powered bot. A Chatbot Corpus containing approx. 290,000 words from 600 simulations of target ChatGPT production was created, representing two GPT versions (3.5 and 4), and three temperature settings. This corpus was then compared with relevant subcorpora in the British National Corpus 2014, which contains 100 million words of British English collected in naturalistic settings. Analyses were conducted at macro- (multidimensional analysis), meso- (comparative frequency analysis), and micro-levels (occurrence of specific pragmatic feature analysis). Findings demonstrated that the ChatGPT-powered chatbot production was systematically more similar to genres of written rather than spoken communication: output demonstrated higher lexical density and was characterised by a relatively low occurrence of features typical of spoken communication such as stance and pragmatic markers. We argue that the methodological framework is applicable across different chatbot models, allowing researchers and developers to use this approach with newer, more refined AI-powered conversational agents in the future.

Introduction

Spoken dialogue systems (SDSs), particularly AI-powered chatbots, have the potential to revolutionise the teaching and assessment of second language speaking through their ability to elicit aspects of interactional competence (IC) that have been difficult to capture in computer-mediated environments. Previous approaches to computer-mediated speaking often relied on semi-direct formats where speakers provide monologic responses to written or pre-recorded prompts (Gablasova et al., 2024; Thirakunkovit et al., 2019). Such tasks constrain the ability to elicit and practice elements of interaction including turn-taking, backchannels, and topic development over multiple turns (Roever & Ikeda, 2022). AI-powered chatbots offer a possible remedy for these problems by simulating a live human interlocutor with the capacity for interaction, the provision of relevant responses, and – if delivered through an avatar – non-verbal cues. AI-powered chatbots thus provide a potential solution to eliciting more authentic spoken communication in tasks delivered on a mass scale (Karatay & Xu, 2025).

However, the use of chatbots powered by Large Language Models (LLMs) such as ChatGPT raises a crucial question: how authentic is the language produced by the chatbot interlocutor? Inauthentic language use in chatbot productions would risk weakening the validity argument for speaking tasks used for assessment and undermining the usefulness of systems designed for pedagogical purposes (Xi, 2023). If chatbot productions fail to replicate the characteristics of talk in target language use domains, their widescale use may lead to negative washback (Voss & Waring, 2025). Evaluating the authenticity of chatbot production, then, is a vital step for the development of such systems; shortcomings must first be identified to allow for fine-tuning in the development process, and to provide robust validity evidence for the use of chatbots as conversational partners. Moreover, efficient evaluation methods are needed to allow for assessment within fast-moving programs of innovation.

In this article, we propose a novel methodological framework for applying corpus linguistic methods to evaluate the authenticity of chatbot production in comparison to (spoken) production in a target domain. The framework comprises three analytical levels – macro, meso, and micro – each offering flexibility to focus on linguistic features relevant to the assessment or pedagogical aims of chatbot use. Specifically, the framework enables comparison of chatbot production with authentic language in a target language use (TLU) domain, and across different chatbot candidates. We illustrate this approach through a case study from the development of a British Council-designed low-stakes formative assessment system – AiBC – in which learners interact with a ChatGPT-powered bot and receive feedback on their performance. Chatbot configurations in the study were defined by two key variables: i) GPT version, which represents different model capabilities and ii) temperature, which affects the probabilistic nature of chatbot output. We discuss how the methodological framework can be used in chatbot development cycles to guide and fine-tune further training, or to evaluate final chatbot productions.

Chatbots and authenticity

AI technology – and particularly generative AI (GenAI) – has been integrated into language teaching and assessment practice at a rapid rate. While aspects of GenAI use in high-stakes language assessments have been critiqued, there has also been recognition that GenAI holds great potential for realising more tailored and personalised formative assessment (Harding, 2025). As Jang and Sawaki (2025) state, advances in GenAI can "facilitate performance processes, engage learners in interactive assessment tasks, provide real-time feedback, and generate personalized, scaffolded language-learning opportunities" (p. 362).

Within the suite of GenAI applications to language education, and formative language assessment more specifically, AI-powered chatbots/conversational agents have been among the

most revolutionary. As noted above, chatbot interlocutors provide scope for more human-like interactions in digitally mediated language learning and assessment environments (Chapelle, 2025; Jeon et al., 2023; Saeki et al., 2024). Research on SDSs has provided empirical support, showing that interactions with a machine interlocutor can elicit some relevant aspects of interactional competence in learner/test-taker productions (Ockey et al., 2023). For example, Timpe Laughlin et al. (2024) found that the syntactic complexity, lexical variety, and fluency of learner productions were broadly comparable in role-plays delivered in both an SDS mode and a face-to-face mode with a human interlocutor. With respect to AI-powered chatbots/conversational agents, Karatay and Xu (2025) found that an AI interlocutor powered by GPT-4o was able to elicit a range of interactional features in test-taker discourse, particularly aspects related to turn-taking and topic development.

These findings suggest that with improving SDS technology, the ability to elicit IC features in pedagogical and assessment contexts will continue to converge on what is currently possible in human-human interactions. However, research to date has typically focused on test-taker/learner productions, with less attention to the productions of chatbots themselves. Such a focus is vital, however, because the *authenticity* of chatbot utterances has the potential to support or undermine validity arguments for the use of systems both in higher- and lower-stakes settings. This point is articulated clearly by Xi (2023), who identified potential "vulnerable" inferential links in a validity argument for speaking assessments involving chatbot interlocutors:

> Limitations of the SDS technology may constrain the types of conversation/writing exchange used, leading to construct under-representation or lack of task authenticity. SDS generated dialogues may lack the kind of diversity and richness of real-world

> conversations and may sometimes be perceived as too contrived compared to naturally occurring conversations.
>
> (Xi, 2023, p. 361)

According to Xi (2023), this "domain definition inference vulnerability" specifically concerns the authenticity of chatbot productions: the extent to which the linguistic characteristics of chatbot utterances mirror those of the relevant target language use domain. This view of authenticity is in line with Bachman's (1991) concept of "situational authenticity", defined as "the perceived relevance of the test method characteristics to the features of a specific target language use situation" (p. 690), which remains an underpinning concept in communicative language teaching and assessment (Hasrol et al., 2022).

In this study, we draw on the notion of situational authenticity and its iterations in Bachman and Palmer's subsequent work (e.g., "real life authenticity" Bachman & Palmer, 1996; "relevance warrant" Bachman 2005) as well as synthesised understandings of authenticity reported by Hasrol et al. (2022) and retool these concepts for a definition of authenticity in chatbot production. Specifically, extending on phrasing used by Bachman (2005, p.18), we define *authenticity* in this context as: "the degree of correspondence (similarity) between characteristics of the language produced by chatbots in a task setting and characteristics of the language produced by human interlocutors in a relevant target language use domain". In this sense, authenticity of chatbot productions is a necessary, though not sufficient, condition of a wider estimation of task/test authenticity which would take into account the full range of test method characteristics.

Given the burgeoning use of chatbots in language learning and assessment, it is vital to establish methods to empirically and rigorously substantiate claims of authenticity. This is particularly important because Xi's (2023) concerns have been confirmed in language

assessment research where SDS productions yielded by specific systems were perceived as "unnatural" (e.g. Ockey et al., 2023, p.392). A wider literature from the field of computer science has also considered the authenticity of chatbot productions as part of a more general quality evaluation, considering whether the chatbot uses "appropriate degrees of formality [and] linguistic register", demonstrates "linguistic accuracy", and provides a "convincing, satisfying, & natural interaction" (Radziwill & Benton, 2017, pp. 6-7). From an educational perspective, exposure to authentic (realistic) representation of (spoken) language during the learning process (e.g. as part of formative assessment delivered by SDSs in which learners practice speaking) is crucial for learners' ability to develop appropriate speaking and conversation skills. In the context of speaking, such exposure prepares L2 learners for the "messiness" of spoken communication (Wagner, 2014) and helps them transfer their knowledge from the practice setting to the real-life environment.

The question of how best to empirically investigate chatbot authenticity, though, remains a live question. Previous studies that have focused on the language used by chatbots in comparison to real-world human interaction highlighted a range of linguistic features related to social and pragmatic aspects of the chatbot production in which chatbots differed from human communication in similar contexts, suggesting useful features to target. For example, Voss and Waring (2025), using a conversation-analytic lens, discussed patterns related to the "naturalness" of the interaction that focused on features such as the deictic and responsive nature of the conversation, which affect the use of linguistic items and strategies (such as the use of personal pronouns and discourse markers). Dippold et al. (2020), drawing on methods and insights from interactional sociolinguistics, conversation analysis and politeness theory, identified issues with expression of alignment and affiliation in chatbot production. Such methods are valuable, but also time consuming, and may not adequately take account of the key characteristics of the target language use domain, a criticism that has been levelled at

studies of authenticity more generally (Hasrol et al., 2022). Thus, in the current study, we turn to corpus linguistics as a means of providing a methodological approach for evaluating chatbot authenticity that is (a) efficient, (b) comprehensive, and (c) deeply connected to the characteristics of the TLU domain through comparison with a relevant reference corpus.

Corpus methods for evaluating chatbot authenticity

Corpus linguistics is a scientific approach grounded in the quantitative analysis of language, drawing on large electronic datasets (corpora) that represent language use in specific linguistic settings (Egbert et al., 2022; McEnery & Hardie, 2011). Pioneered in the 1960s, corpus methods have been central in providing evidence-based description of linguistic features characteristic of discourse across different modes, genres, and registers of communication, and in explaining their relationship to specific communicative goals (Biber et al., 1999; McEnery & Brezina, 2022). Comparative analysis has been integral to this work and several corpus methods, such as multi-dimensional analysis and keyword analysis, have been developed specifically to identify similarities/differences across text types, further supported by foundational corpus techniques such as frequency, concordance and collocation analyses (Brezina, 2018). These methods rely on information about frequency and distribution (dispersion) of target linguistic features to establish the probability that a particular linguistic feature will occur in a text type. Corpus approaches have been applied across different areas of language education and assessment, where corpus-based descriptions of real-life (authentic) language use have informed the development and validation of teaching and testing resources (see Gablasova, 2021; Jablonkai & Csomay, 2022).

To date, corpus research has played a major role in identifying and describing characteristics of spoken production. Speaking constitutes a complex skill which involves an interplay of social, cognitive and linguistic knowledge to produce and process spoken language

(Hughes & Reed, 2016). As a result, a number of linguistic features – grammatical, lexical and pragmatic – are particularly prominent in spoken (interactive) communication. Corpus-based investigations provided a comprehensive account of these features (see Biber et al., 1999; Biber & Conrad, 2019; Caines et al., 2016 for an overview), highlighting elements such as personal pronouns, stance markers and specific grammatical structures (e.g., the progressive aspect). Corpus research has also pointed to the absence or infrequent occurrence of linguistic features typical of other modes/genres as a defining characteristic of speech (Biber et al., 1999). For example, compared with speech, planned writing – such as academic and journalistic prose and writing for TV shows – typically exhibits higher lexical density, higher frequency of passives and nominalisations, with less lexical repetition (Biber & Conrad, 2019).

Although text corpora have traditionally served as key data sources for training the LLMs that support current AI tools (e.g., Dam et al., 2024), corpus-based methods have not yet been systematically applied to evaluate the language production of chatbot technology. We argue, however, that corpus linguistics provides a vital toolkit for developers of assessment and learning systems where chatbots are used for spoken or written interaction. Specifically, corpus linguistics provides methods for macro-level analysis, meso-level analysis, and micro-level analysis, allowing for a multi-layered understanding of chatbot productions that can be generated relatively efficiently (compared to discourse analytic approaches such as Conversation Analysis; CA) and which can capture patterns in large datasets. We illustrate this framework in Figure 1.

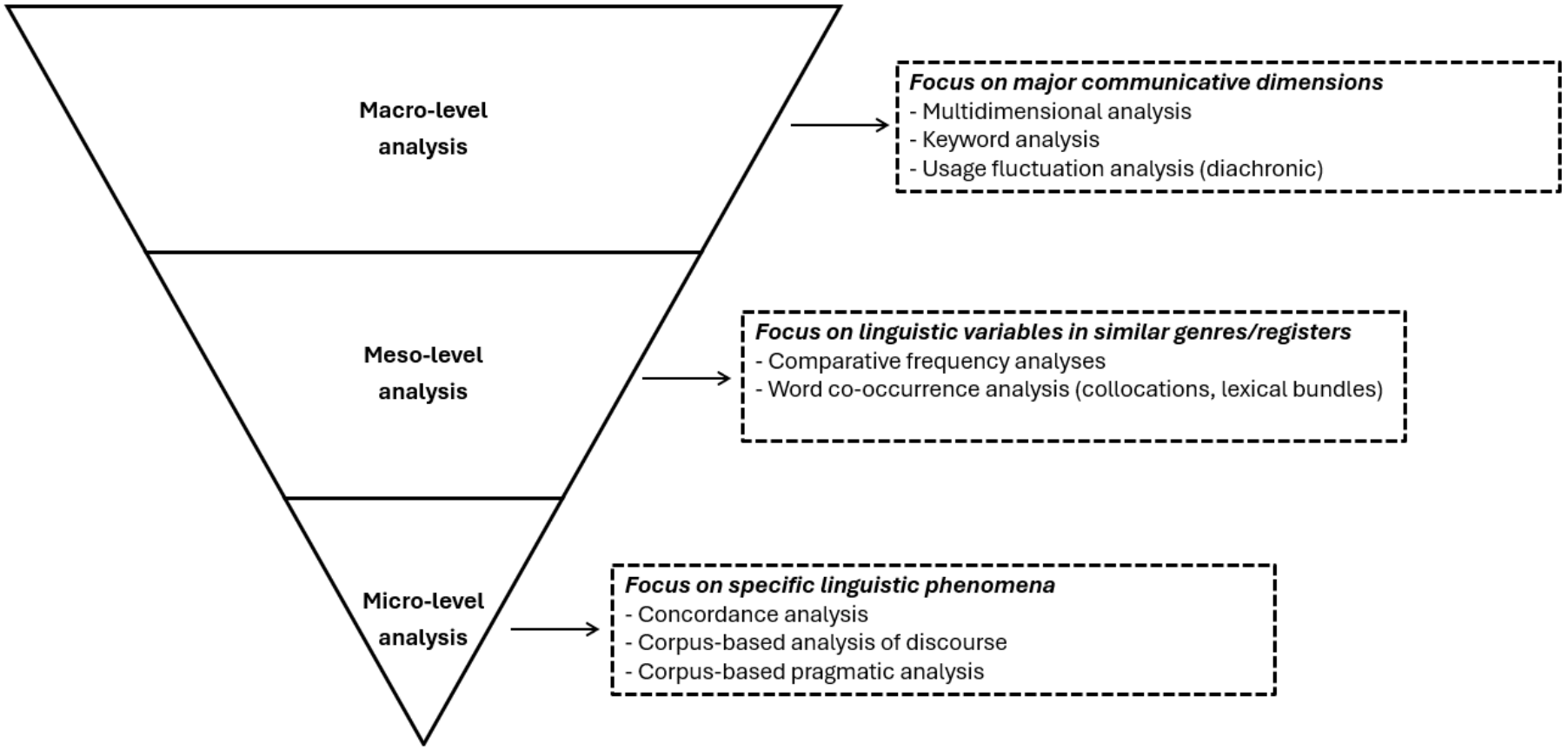

Figure 1. A multi-level corpus-based framework for evaluating chatbot authenticity

Metaphorically, this multi-level approach can be envisioned as progressively zooming in on a map: first, contrasting the global contours of continents (macro level), then identifying the boundaries and characteristics of individual countries or regions (meso level) and ultimately observing the intricate details of specific cities, towns, rivers and mountains (micro level). The framework can be applied flexibly, with each level offering a different perspective on the data, to support a comprehensive understanding of language use in chatbot communication. Depending on the goals of the chatbot development and validation, each level can focus on a distinct set of linguistic features and can be applied independently. Alternatively, the levels can be used cumulatively, with insights from one level guiding the analysis at the subsequent levels. In this article, we illustrate each level of the framework through an analysis of data from the development cycle of a real AI-powered immersive learning and assessment system (AiBC).

Research aims

This study demonstrates how corpus methods can be applied to evaluate chatbot language production during the development and/or validation of SDSs for language learning and assessment. Specifically, we address the following research questions:

RQ1: To what extent are different chatbot versions similar to or different from each other?
RQ2: To what extent are linguistic features of chatbot production similar to or different from authentic spoken communication?

RQ1 investigates the predictability of linguistic variation in chatbot production as determined by specific chatbot versions and settings. Although there exists documentation on these variables (for example, see the description of the temperature setting below), it is currently not possible to predict their effect on specific linguistic features without empirical analysis. RQ1

thus represents a preliminary question, which examines the extent and direction of variation within the chatbots. From the practical perspective, addressing this question is particularly relevant when SDS/chatbot developers wish to evaluate the performance of multiple candidate chatbots in terms of their alignment with the intended communicative purposes of the bots. Building on this foundation, RQ2 directly evaluates the performance of the chatbots with reference to the target domain of communication (interactive speech). This analysis includes a comparison with a broader range of genres varying in interactiveness and formality to provide a comprehensive understanding of authenticity and appropriateness of chatbot performance.

Methodology

The AiBC program

Data in the study were drawn from early-stage trialling during the development of a British Council-designed low-stakes formative assessment system – AiBC - which consists of a series of immersive tasks in which learners interact with a ChatGPT-powered bot. AiBC is designed as an integrated learning and assessment ecosystem which incorporates a spoken dialogue system to replicate real-world workplace tasks. AiBC's primary purpose is to provide adult L2 learners with opportunities to practise and demonstrate their spoken production and interaction skills. Learners engage in an immersive task, interacting with the ChatGPT-based chatbot to achieve a specific goal.

The prototype tasks used for the present study were aimed at learners at the B1 and B2 levels of the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2020), with a focus on oral interaction. Each task was devised predominantly using the high-level scale of *Goal-oriented cooperation* as the main construct. The tasks reside on the British Council English Online platform, a digital learning environment for adult learners (https://englishonline.britishcouncil.org/ai-speaking-practice/). The focus, topic, and level of

the speaking tasks were aligned with modules chosen from the target levels within the English Online course. The tasks are low stakes, providing formative feedback for use by students and teachers, highlighting areas for improvement in spoken performance within the course context, and supplementing existing teacher-mediated speaking assessments. Task results are not used for certification or for high-stakes decisions outside of the English Online course.

Task development for AiBC has drawn heavily on the Sociocognitive Model (Chalhoub-Deville & O'Sullivan, 2020), the Comprehensive Learning System (O'Sullivan, 2021) and the CEFR. In terms of task design, the tasks were created to offer an integrated-skills, multimodal experience, combining all modes of communication in the CEFR – reception, production, interaction and mediation – using textual and non-textual visual and audio input. The output is expected to contain features of productive and especially interactive language. Although each task has a clear structure and end goal, it is largely unscripted, with machine interlocutor responses generated through ChatGPT working from a prompt including the transcript of the dialogue thus far. The total time required to complete each task is up to 25 minutes.

At the beginning of the task, the learner is presented with the purpose of the task, for example, selecting and justifying a new strategy to increase sales at a company. They are then given material to help them make a decision (information sheets of relevant facts; statistics and video input explaining current strategy). After preparation time, the learner begins the dialogue with the chatbot, giving their response to the assigned task using the input materials provided. The interaction is designed to represent a real-world scenario, such as a workplace meeting. The role of the chatbot as the interlocutor is to carry the learner through the task, asking questions which can guide weaker performing learners while challenging stronger learners to justify their ideas to ensure that the task constructs are elicited and the conversation stays on track. When sufficient opportunity has been given to achieve the task goal, the chatbot brings

the conversation to an appropriate conclusion and ends the task, with the learner receiving task-specific feedback on their performance.

In terms of training, the performance of the AiBC chatbot was guided by a carefully designed prompt developed over several testing cycles. The prompt that generated the chatbot productions used in the case study was drawn from an early stage of the training round. The prompt provided a detailed set of instructions regarding key characteristics of the chatbot's communicative behaviour which addressed the following elements: i) Personal characteristics of the speaker represented by the bot, with guidance on personality attributes (e.g. supportive and friendly) and internal states of mind; ii) the communicative role of the chatbot, defined in terms of the professional role represented in the dialogue (e.g. a company representative, line manager) and specifying relevant areas of expertise and prior knowledge; iii) interactional behaviour, with extensive guidance on the types of questions and responses, communicative aims of the conversation, discourse management strategies (such as openings, closings and topic transitions), the degree of interactiveness, and typical turn length, limited to thirty-five words per turn to avoid long or repetitive outputs known to occur in chatbot interaction (e.g., Dippold et al, 2020); and iv) features of language use related to style and socially appropriate language; for example, the bot was instructed (through an extensive prompt) to produce language typical of spoken communication of young speakers of British English, to listen and ask questions, and to maintain friendly tone with a professional but casual style (this is a summary of a more detailed set of guidelines). Overall, the instructions addressed the desired features, pre-empting known issues in chatbot-generated production (e.g. lengthy responses, limited interactiveness or tendency for factual inaccuracies) and addressing any similar, unanticipated bot-related issues that arose during testing.

The AiBC system was trialled over multiple testing cycles, involving both human-generated and automated evaluation, to assess the performance of the task design, materials

and task prompt. This was achieved by means of a purpose-built development tool, which produced text-based simulations of a task dialogue to be reviewed by the development team as part of internal testing cycles. The present study draws on data generated through this process during the early stages of the development.

Independent variables: GPT version and temperature

The model type and temperature setting are two key design variables to be considered in chatbot development, as they have a significant impact on shaping chatbot output in relation to the intended purpose of the system. Currently, limited information is available – either in product documentation (OpenAI, 2024a, OpenAI, 2024b, OpenAI, 2024c) or based on research to date – on how these variables may affect various linguistic features in chatbot-generated language.

ChatGPT (Generative Pre-trained Transformer) is an AI chatbot which interacts with users via writing, audio chat, or an automated voice system (Advanced Voice Model) created by OpenAI (2024b). It simulates human communication and engages in a dialogue with the user, responds to iterative instructions, and challenges incorrect premises (OpenAI, 2024b).

To examine the differences in chatbot linguistic output, two ChatGPT versions were investigated in this study - GPT-3.5 and GPT-4. GPT-3.5, the first iteration of ChatGPT, finished training in 2022; ChatGPT-4 was launched in 2023 with advanced reasoning abilities, the capability to interpret visual input and longer texts, and improved content generation features (OpenAI, 2024c). OpenAI noted that,

> In a casual conversation, the distinction between GPT-3.5 and GPT-4 can be subtle. The difference comes out when the complexity of the task reaches a sufficient threshold—GPT-4 is more reliable, creative, and able to handle much more nuanced instructions than GPT-3.5. (OpenAI, 2024c)

This suggests that the version-related variation in bot language could be relatively minor in conversational tasks; however, in practice, it is difficult to predict how each version responds to the same prompt, and its impact on specific linguistic features in chatbot production (Lahat et al., 2024; Lee, 2023).

Temperature is another variable with an impact on linguistic patterns in bot production, affecting the degree of randomness with which the next word is selected in the output. ChatGPT allows temperature setting ranging from 0 to 2, stating that higher values "will make the output more random" and lower values "will make it more focused and deterministic" (OpenAI, 2024a). The lower settings are thus expected to result in more predictable and "consistent outputs" (Davis et al., 2024, p. 1), with higher temperature leading to increased creativity and diversity (Peeperkorn et al., 2024; Renze & Guven, 2024). However, the effect of the variable on (linguistic) output is currently not well understood, with a very limited number of studies that investigate this variable systematically (Li et al., 2025, p. 243). As a result, as Renze and Guven (2024, p. 7374) argue, the current practice of choosing a temperature setting "is largely based on guesswork, gut instinct, non-systematic experimentation, and iterative refinement". It is thus difficult to predict any systematic impact on temperature settings on GPT outputs in different tasks.

Therefore, to investigate the effect of temperature, this study examined linguistic patterns at three levels (0, 0.5 and 1) in order to represent a broad range of values and to allow for a direct comparison between the two GPT versions (GPT-3.5 works with the range of 0-1 temperature, while GPT-4 range was expanded to 0-2). The setting above 1 was avoided as a higher temperature was argued to result in more hallucinations or non-sensical responses (Lee, 2023).

Given the complexity of LLM training and functionality, different chatbot versions are highly sensitive to even minor variation in input phrasing, which can result in unpredictable variation in their language output (Knoth et al., 2024). This factor makes it essential to systematically evaluate the output of AI-powered systems, as their communicative behaviour cannot be predicted from documentation or prompts alone.

Corpus design and composition

The Chatbot Corpus

The Chatbot Corpus (Hazelhurst et al., 2025) was used for the analysis. Drawing on the selected independent variables, six GPT subcorpora were generated using the same prompt, across two GPT versions (3.5 and 4) and three temperature settings (0, 0.5 and 1). For each temperature setting and GPT version, 100 simulations were created, resulting in a total of 600 simulations. Each simulation consisted of five turns with another bot serving as the second interlocutor. Throughout the study, the following naming convention is used to refer to the different chatbot subcorpora: 'CH' (Chatbot version), followed by the GPT version (3.5 or 4) and the temperature setting (0, 0.5 or 1). Thus, for example, the data representing GPT-4 with a temperature setting of 0.5 would be referred to as CH4(0.5). Table 1 and Figure 2 provide an overview of the Chatbot Corpus composition in terms of size per each variable combination.

Table 1 Chatbot Corpus composition and size

GPT version	Temp. setting	Average length per simulation (tokens)	SD	Subcorpus size (tokens)
3.5	0	185.2	7.5	18,521
	0.5	184.1	21.7	18,408
	1	183.3	27.3	18,334
4	0	106.0	7.0	10,599
	0.5	145.5	43.5	14,550

1	108.2	12.5	10,825

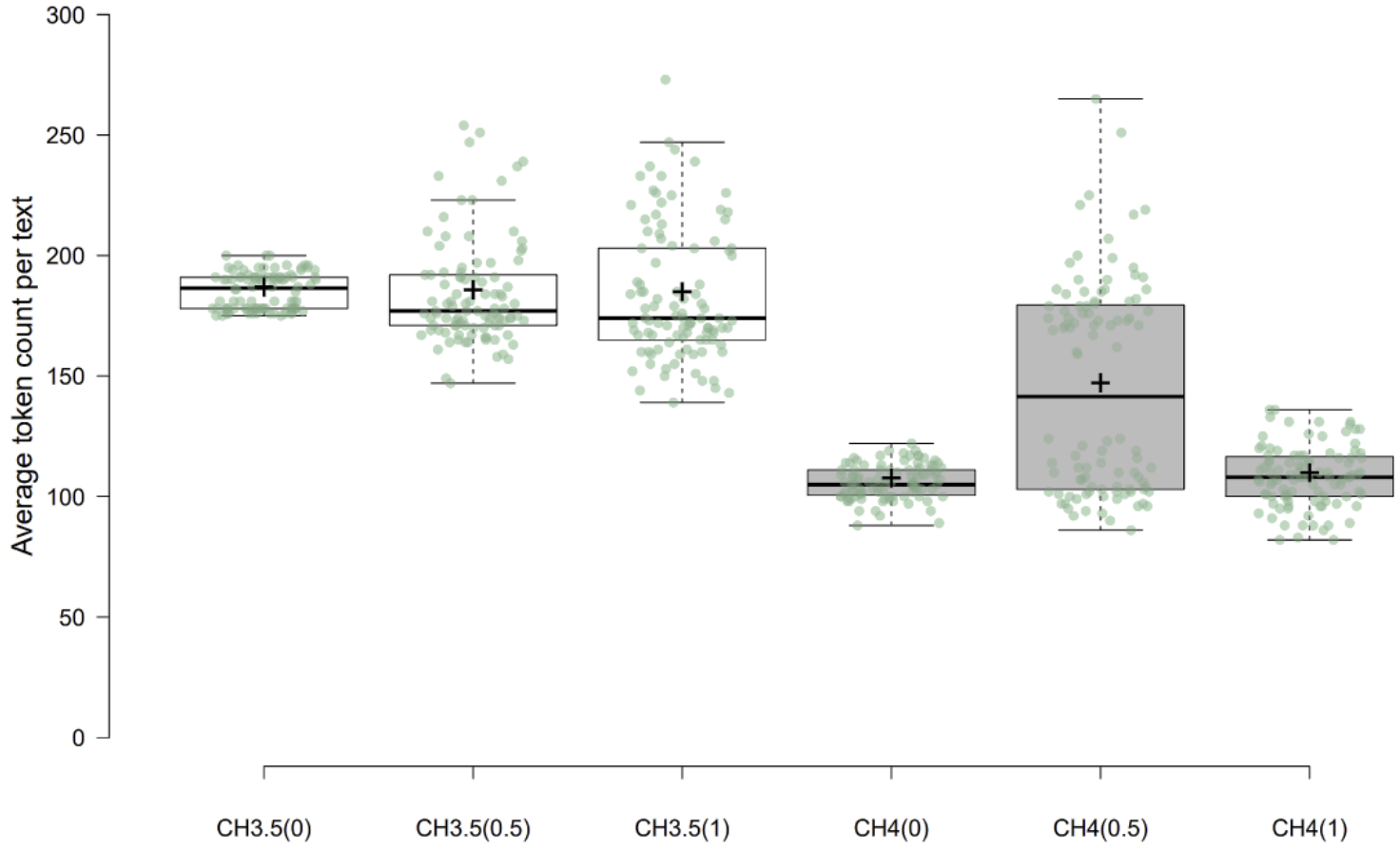

Figure 2 Average token count for each Chatbot subcorpus [X-axis label format: GPT (3.5/4), Temperature (0/0.5/1)]

As Table 1 and Figure 1 show, both the GPT version and temperature affected the length of the chatbot output, with GPT-3.5 simulations producing longer turns than GPT-4. The effect of temperature on word count was more pronounced in GPT-4 than GPT-3.5, with CH4(0.5) showing the greatest variation in output length.

BNC2014

The British National Corpus 2014 (BNC2014; Brezina et al., 2021; Love et al., 2017) was used as a reference corpus representing current British English. It comprises over 100M words from 88,171 text files from different spoken and written genres. All major genres (see Table 2) were used in multi-dimensional analysis, with a subset further used in comparative frequency analysis. The 'Informal speech' subcorpus represents spoken interaction between acquaintances, friends and family members. 'Written-to-be-spoken' texts include scripted language (e.g., drama and TV shows) intended to reflect spoken interaction. 'Elanguage' includes online texts such as blogs, social media posts, and e-mails, representing informal or semi-formal writing. 'Newspapers' and 'Magazines' consist of formal and semi-formal journalistic prose, while formal writing is represented by 'Academic prose' (academic books and journals), and 'Official documents' (e.g., reports). Table 2 provides an overview of the subcorpora according to text and token size.

Table 2 BNC2014 subcorpora according to their size

BNC2014 subcorpus	No of texts	Size (no of tokens)
Academic prose	2,879	1,970,1027
Elanguage	2,381	529,1594
Fiction	1,069	2,043,2736
Informal speech	1,251	11,029,483
Magazines	26,965	15,297,596

Newspapers	50,210	20,338,500
Official documents	2,690	7,040,145
Written-to-be spoken	726	3,174,165

Analytic procedure

Applying the framework

As noted above, corpus linguistics represents a powerful methodological approach for quantitative language analysis (McEnery & Brezina, 2022), providing a flexible and efficient means of evaluating chatbot authenticity across multiple levels of analysis. In this study, we employed methods from each analytical level from the framework introduced earlier: i) At the **macro** level, we conducted Multi-Dimensional Analysis (MDA; Biber, 1988), a statistical method used to compare and classify language use across genres of speech and writing. MDA employs factor analysis to identify systematic variation in the co-occurrence of lexical and grammatical features associated with different communicative functions (dimensions). It enables a comprehensive synthesis of linguistic variation by showing how text types place along major communicative dimensions. The features entered into MDA are functionally motivated, selected for their relevance to communicative goals and the research focus. In this study, MDA offered a comparative overview of chatbot language relative to established genres/registers from the BNC2014 reference corpus, providing insights into chatbot authenticity at a macro level. ii) Drawing on findings from MDA, at the **meso** level, we conducted comparative frequency analysis to examine similarity in the use of a set of target linguistic variables in chatbot and authentic human production. Compared to the MDA, which highlights clusters of linguistic features that systematically co-occur, comparative frequency analysis allows a more fine-grained and focused investigation of the linguistic behaviour (e.g., frequency and distribution) of individual features relevant to the construct of speaking in the chatbot development project. iii) Finally, at the **micro** level, we analysed the realization of a specific pragmatic function (a response to an apology) in the Chatbot Corpus, comparing it

with language produced for a corresponding function in the BNC2014 to illustrate how authenticity can be investigated through corpus techniques at the utterance level. The micro level offers the opportunity to focus on a (highly) specific linguistic feature and to analyse its use at the level of discourse, to develop finer-grained insights into appropriateness of use.

To assess the similarity of chatbot output to human production drawn from the target language use domain, at each level of the analysis chatbot production was compared to both target (informal speech) and non-target genres (e.g. written-to-spoken texts, elanguage, newspapers or academic prose). The comparison with target genres allows for an assessment of the bot's alignment with desired communicative patterns; the inclusion of non-target genres can reveal systematic occurrence of linguistic features (e.g. typical of writing) that may run counter to training intentions and may require further attention during development. This dual approach provides a comprehensive understanding of the nature of chatbot language outputs, demonstrating the versatility of corpus methods for evaluating different aspects of chatbot production.

Feature selection for macro-, meso- and micro-level analyses

Drawing on previous corpus research on spoken production (e.g. Caines et al., 2016; Biber & Conrad, 2019), three groups of linguistic features were analysed to establish the similarity between chatbot production and the target genre of spoken informal communication[1]: i) *MDA* included a broad range of syntactic, lexical and pragmatic features (Biber, 1988) such as negation, passive voice, nominalisation, pragmatic markers and pronouns. This allowed a comprehensive categorisation of chatbot language along a cline of established spoken and written genres with particular emphasis on features characteristic of spoken interaction (e.g., pragmatic and stance markers). The full list of features analysed in the MDA is available in the

[1] It should be noted that a more comprehensive set of features was analysed during the development and internal testing of the AiBC system. Only a selected set of features is reported in this study to illustrate how the framework can be applied.

Supplementary Appendices. ii) *Comparative frequency analysis* focused on the linguistic features highlighted by MDA and strongly associated with interactive speech, as shown in previous corpus-based research on speaking assessment (Gablasova et al., 2024). These included: *Lexical density* as a measure of informational density determined by the proportion of content words (nouns, verbs, adjectives, adverbs) to total words (Brezina, 2018). It is expressed on a scale of 0-1, with higher numbers indicating greater density; *Epistemic stance markers* such as *I think, maybe* and *certainly* which fulfil multiple functions in social interaction and meaning negotiation (Kärkkäinen, 2003); *Spoken pragmatic markers,* such as *ah, alright, okay, oh* and *well,* which help manage the flow of discourse, acknowledge the receipt of information or indicate a topic change (Aijmer, 2003); and *Amplifiers* and *downtoners,* such as *absolutely, fully, almost,* and *barely,* which are used to indicate degree on a scale (Biber et al., 1999). The full list of searched forms is available in the Supplementary Appendices. iii) The micro-level frequency analysis focused on the forms used to express a specific speech act typical of spoken interaction – a response to an interlocutor's mistake and subsequent acknowledgment or apology. This analysis examined whether the chatbots produced authentic responses to speech acts and language functions commonly found in conversations (e.g., apologies, requests and questions). This particular speech act was selected as an illustrative example for this part of the analysis.

Corpus and statistical analyses

The data were analysed using #LancsBox X (Brezina & Platt, 2024), a free software package for corpus and statistical analysis. For the macro-level analysis, data were processed using the Wizard tool, which allowed batch searching for a large number of linguistic variables in individual texts of the corpora, producing a complex frequency matrix for MDA. The complete set of search terms is available in the Supplementary Appendices. For the meso and micro levels, the Text tool was used to search for individual forms and their distribution in the target

and the reference corpora and to note their frequency per each text. The absolute frequency was normalised to 1,000 words, to allow for comparison across sub-corpora. The data were further processed for each analysis using the R package (R Core Team, 2025) with custom-built R scripts available in #LancsBox X.

To answer RQ1 and RQ2, different statistical tests were carried out. At the macro-level, the MDA was conducted following the standard procedure (Biber, 1988; Brezina, 2018). At the meso-level, inferential statistics (Welch's ANOVA) were carried out, with results visualized via boxplots and 95% confidence intervals. As Levene's tests for equal variances across groups showed unequal variances, Welch's ANOVA, with adjusted degrees of freedom, was used as an omnibus measure, with Bonferroni-adjusted comparisons (independent t-tests) used to show statistically significant differences across individual chatbots and genres. At the micro-level, a comparative analysis was performed to establish if the target structure has been attested in human production (as represented by the BNC2014).

Results

This section first presents the results of the more comprehensive MDAs, which in turn guided the selection of target linguistic features for further analyses.

Macro analysis: MDA

MDA identified three dimensions (functional scales) visualised in Figures 3a-c and described below.

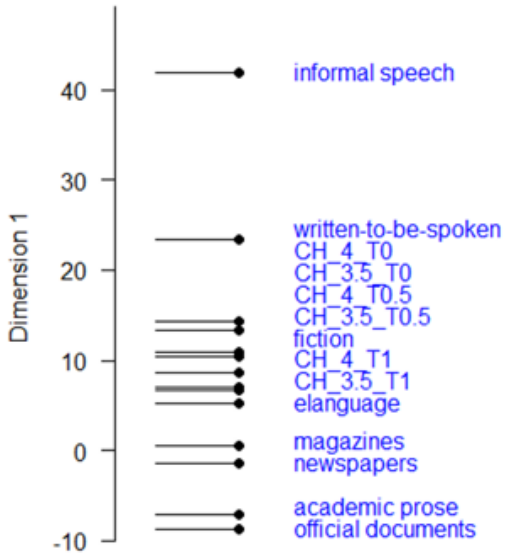

Figure 3a Dimension 1: Involved-informational

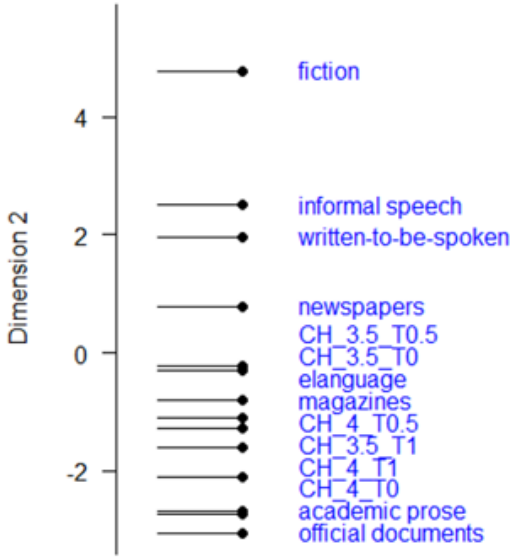

Figure 3b Dimension 2: Narrative-descriptive

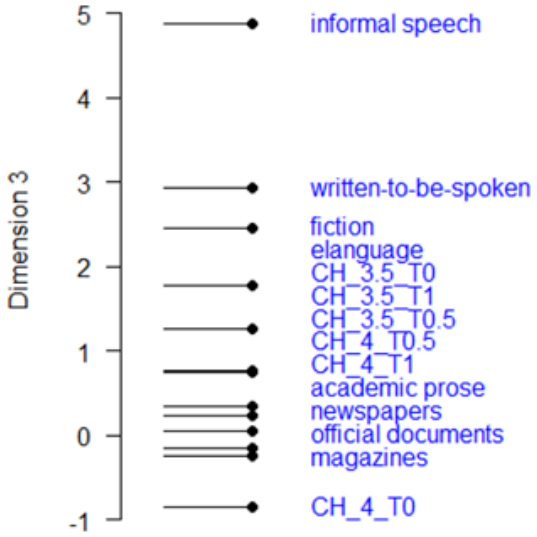

Figure 3c Dimension 3: Infinitival- informationally-dense

Dimension 1

The first dimension identified in MDA aligned strongly with Biber's *involved–informational* dimension, with several features pointing clearly towards involved, spoken discourse on the positive end. Contractions (factor loading of 0.849) were the strongest positive contributor, followed by present tense verbs (0.701) and second person pronouns (0.533), reflecting the real-time and interpersonal nature of spoken interaction. This interpretation is further reinforced by features such as first-person pronouns (0.476), epistemic verbs (0.502), analytical negation (0.505), and spoken discourse markers like *you know* (0.459) and *so/well* in initial position (0.510). These forms are characteristic of texts that emphasize personal stance, engagement, and speaker presence, such as casual conversations or interviews. The negative side of this dimension was defined by features indicative of informational, literate discourse such as prepositions (-0.653), nominalizations (-0.524), and passive constructions (-0.442), typically associated with dense, abstract written texts such as academic articles or official reports.

The reference genres loaded onto the dimension in the expected order (see Figure 3a), ranging from informal speech, written-to-be-spoken texts, fiction, and elanguage—all aligning with the involved end of the spectrum; newspapers, magazines, academic writing, and official documents clustered closer to the informational end. Productions of the different chatbots (labelled by version and temperature setting: CH_[version]_[temperature setting]) fell in the middle, between written-to-be-spoken texts and elanguage. Interestingly, the temperature setting, rather than the chatbot version, appeared to influence the placement on this dimension, with outputs at temperature 0 aligning more closely with *involved* discourse, and those at temperature 1 shifting towards the *informational* end.

Dimension 2

Dimension 2, labelled *narrative–descriptive*, distinguished between storytelling and descriptive styles of discourse. The strongest positive loading came from past tense verbs (1.019), a clear marker of narrative discourse, with events typically recounted in the past. This was supported by a moderate contribution from third person pronouns (0.480), which often feature in narratives involving characters or external agents. Additionally, public verbs (0.404), such as *said* or *reported,* further reinforced the narrative framing. In contrast, adjectives (-0.359) showed a negative loading, suggesting that descriptive texts, which rely heavily on modifiers and detail, lie on the opposite end of this dimension.

Along this dimension, Figure 3b shows the chatbots clustered towards the *narrative* side, between newspapers and academic prose, and alongside elanguage and magazines. This suggests a tendency toward more descriptive or expository styles. Unlike Dimension 1, where temperature played a greater role, here it was primarily the chatbot version that influenced positioning. Versions 4 and 3.5 clustered closely together, indicating similar discourse tendencies, while CH3.5(1) stood as a notable exception.

Dimension 3

Dimension 3 – *infinitival–informationally-dense* – explained the least amount of variation in the dataset and was defined by only three variables with factor loadings above 0.3, suggesting a relatively narrow functional scope. The strongest positive contributor was the infinitive marker (0.329), indicating the presence of non-finite verb constructions, while the strongest negative loadings came from other noun types (-0.919) and lexical density (-0.552). This dimension appears to contrast texts that rely more on non-finite, abstract, or procedural constructions—such as those found in instructions or objective descriptions—with those that are noun-heavy and lexically dense, characteristic of more technical or information-loaded

prose. Although its explanatory power is limited, this dimension may capture subtle stylistic differences in clause structure and syntactic complexity across genres/chatbot outputs.

Chatbots were relatively widely dispersed along Dimension 3 (see Figure 3c), with CH4(0) representing the content-heavy end, aligning more closely with magazines and official documents. The overall clustering along this dimension appeared to be more influenced by the version of the chatbot rather than by the temperature, indicating that structural and syntactic choices—particularly around lexical density and noun usage—are more closely related to the model architecture than other configurations such as temperature.

Meso analysis: Target linguistic features

Lexical density

A Welch's ANOVA revealed a statistically significant effect of the text type/chatbot on the degree of lexical density [$F(9.0, 382.5)=312.48$, $p<.001$, $\omega 2 = 0.58$], with Bonferroni post-hoc comparisons reported in Table 3. As shown in Figures 4a-b, lexical density was typically higher in written genres such as academic prose ($M=0.60$, $SD=0.03$) and elanguage ($M=0.58$, $SD=0.03$), followed by written-to-be-spoken production ($M=0.55$, $SD=0.02$) and informal conversations ($M=0.50$, $SD=0.02$).

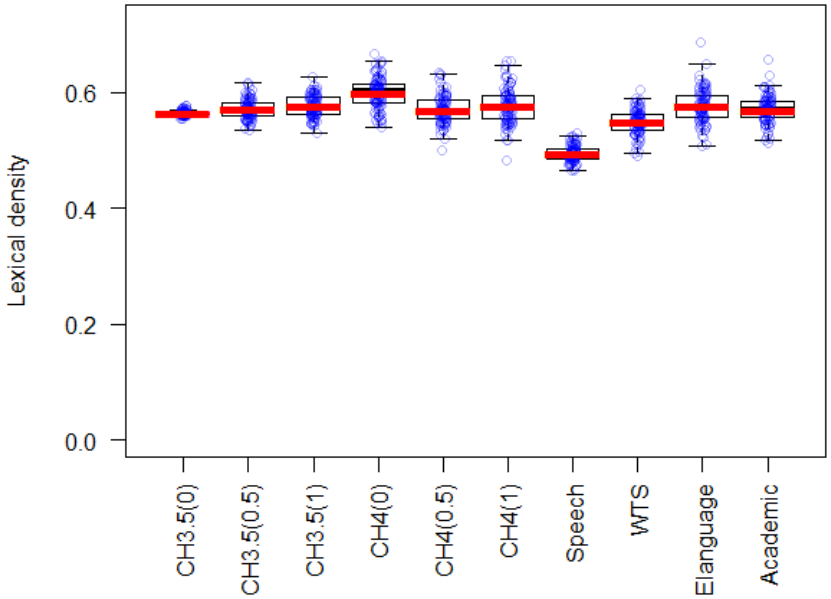

Figure 4a Lexical density: Boxplots with individual data points overlaid

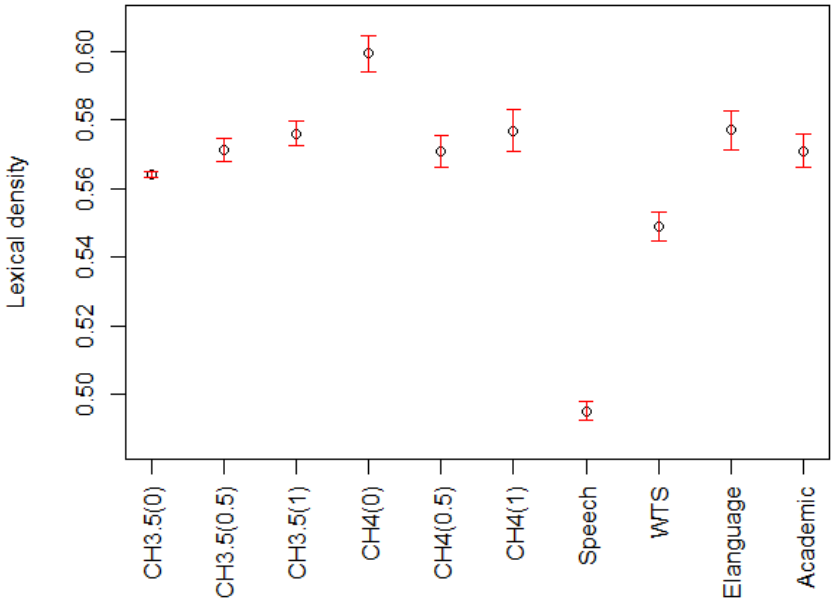

Figure 4b Lexical density: Means with 95% confidence intervals

Regarding the lexical density of chatbot productions, two of the chatbots – CH3.5(0) and CH4(0) – differed from all other chatbots; however, the same temperature setting (0) resulted in the opposite effect in these two chatbots, reducing the density in CH3.5(0) (*M*=0.56, *SD*=0.004) while increasing it in CH4(0) (*M*=0.60, *SD*=0.03). No statistically significant differences were observed between the output of the remaining four chatbots – CH3.5(0.5), CH3.5(1), CH4(0.5) and CH4(1). When compared to BNC2014 genres, the lexical density of chatbot output across all GPT versions and temperature settings was considerably higher than that in authentic informal speech, and somewhat higher than written-to-be-spoken production, aligning instead with academic prose and elanguage. No statistically significant differences were found between four chatbots – CH3.5(0.5), CH3.5(1), CH4(0.5) and CH4(1) – and elanguage and academic prose; there was a statistically significant difference between CH4(0) and both of these genres, with CH4(0) showing an even higher lexical density.

Table 3 Lexical density: Post-hoc Bonferroni comparisons

	CH3.5(0)	**CH3.5(0.5)**	**CH3.5(1)**	**CH4(0)**	**CH4(0.5)**	**CH4(1)**
CH3.5(0.5)	.002**					
CH3.5(1)	.001**	1.00				
CH4(0)	.001**	.001**	.001**			
CH4(0.5)	.307	1.00	1.00	.001**		
CH4(1)	.003**	1.00	1.00	.001**	1.00	
ACAD	.222	1.00	1.00	.001**	1.00	1.00
ELAN	.001**	1.00	1.00	.001**	1.00	1.00
WTS	.001**	.001**	.001**	.001**	.001**	.001**
Speech	.001**	.001**	.001**	.001**	.001**	.001**

Epistemic stance markers (ESMs)

A Welch's ANOVA found a statistically significant effect of the text type/chatbot on ESM frequency [$F(9, 558.32)=1336.1$, $p<.001$, $\omega 2 = 0.61$], with post-hoc Bonferroni comparisons shown in Table 4. Figures 5a-b show that ESMs are a very frequent feature in spoken interactive discourse (*M*=15.59 occurrences per 1,000 words, *SD*=4.96), which is also indicated by their

higher frequency in the written-to-be-spoken genre (*M*=8.28, *SD*=3.57). They are less frequent in elanguage (*M*=5.39, *SD*=4.54) and relatively rare in academic discourse (*M*=2.38, *SD*=1.74). In terms of ChatGPT versions, GPT3.5 seemed to include ESMs at a consistent rate across all temperature settings (with means ranging from 4.63 to 5.29) while GPT4 produced nearly no ESMs at 0 temperature (*M*=0.1, *SD*=0.97), with somewhat higher occurrence at 0.5 (*M*=2.68, *SD*=3.22) and 1 (*M*=2.20, *SD*=4.90). There was no statistically significant difference between the three GPT3.5 versions, all of which differed from the GPT4 versions. Regarding the GPT4 versions, there was no difference between CH4(0.5) and CH4(1), both of which differed from CH4(0).

Compared to BNC2014, chatbots across all GPT versions and temperature settings produced ESMs at a considerably lower rate than informal speech and written-to-be-spoken texts – this difference was statistically significant. Chatbot production was generally most similar to academic prose and elanguage, with the exception of CH4(0) which produced even fewer ESMs than was found in these two written genres.

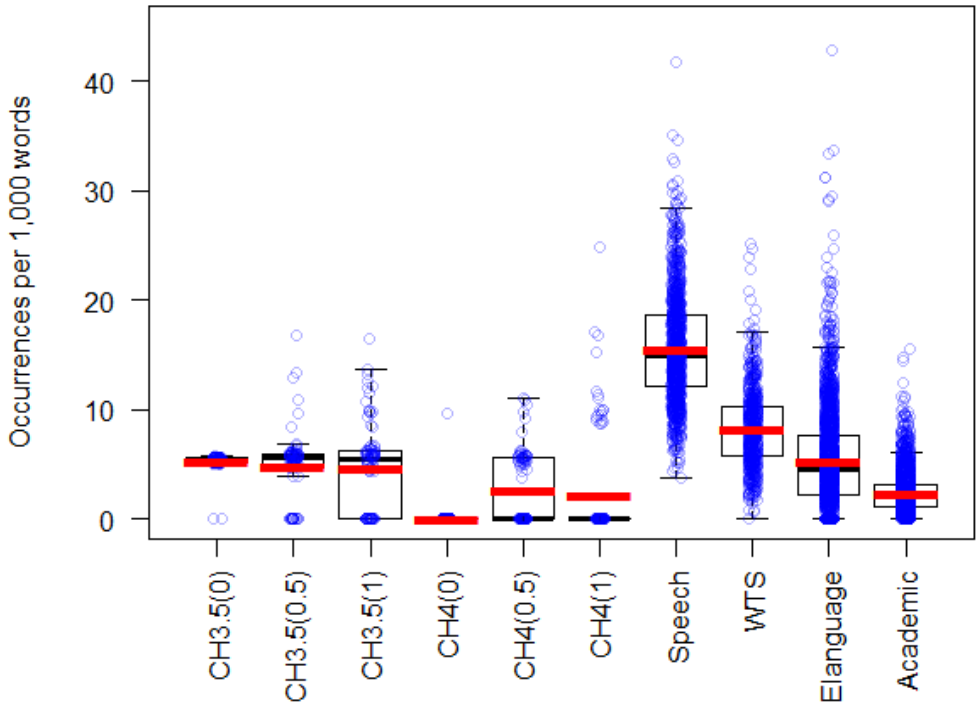

Figure 5a Stance markers: Boxplots with individual data points overlaid

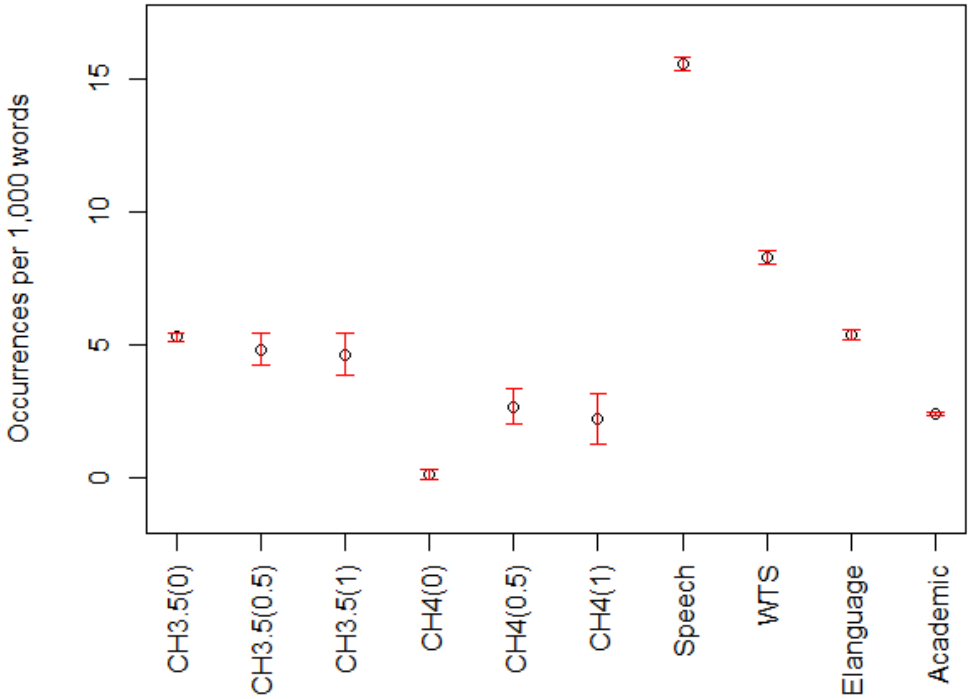

Figure 5b Stance markers: Means with 95% confidence intervals

Table 4 ESMs: Post-hoc Bonferroni comparisons

	CH3.5(0)	CH3.5(0.5)	CH3.5(1)	CH4(0)	CH4(0.5)	CH4(1)
CH3.5(0.5)	1.00					
CH3.5(1)	1.00	1.00				
CH4(0)	.001**	.001**	.001**			
CH4(0.5)	.001**	.001**	.008**	.001**		
CH4(1)	.001**	.001**	.007**	.002**	1.00	
ACAD	.001**	.001**	.001**	.001**	1.00	1.00
ELAN	1.00	1.00	1.00	.001**	.001**	.001**
WTS	.001**	.001**	.001**	.001**	.001**	.001**
Speech	.001**	.001**	.001**	.001**	.001**	.001**

Spoken pragmatic markers (SPMs)

A Welch's ANOVA showed a statistically significant effect of the text type/chatbot on the SPM frequency [$F(10.0, 1173.1)=2{,}273$, $p<.001$, $\omega 2 = 0.797$]; Bonferroni post-hoc comparisons are provided in Table 5. As seen from Figures 6a-b, similar to ESMs, SPMs represent a very typical feature of informal spoken interaction (*M*=49.97 per 1,000 words, *SD*=14.11) with a somewhat lower occurrence in written-to-be-spoken texts (*M*=14.33, *SD*=8.87). These markers play a very small role in written genres such as elanguage (*M*=1.84, *SD*=3.11) and academic prose (*M*=0.28, *SD*= 0.81). Regarding chatbot production, GPT3.5 chatbots contained no [CH3.5(0) and CH3.5(0.5)] or hardly any markers [CH3.5(1)]. In GPT4, the occurrence of SPMs increased with temperature from no occurrence in CH4(0) to 4.3 occurrences per 1,000 words in CH4(1) (*M*=4.27, *SD*=6.75). A closer analysis revealed that the SPM increase in GPT-4 can be mainly attributed to the use of two markers (*so* and *alright*) with *so* accounting for over half of all occurrences. *Ah*, *oh*, *okay*, *right* and *yeah* all occurred in fewer than ten instances. Overall, all chatbots produced no or very few SPMs, with the exception of CH4(1), which differed statistically from the remaining bots.

When compared to BNC2014, chatbot production was close to academic prose, followed by elanguage. Only CH4(1) produced more SPMs and was thus closer to written-to-

be-spoken texts. However, as shown by Figure 6b, the difference between real-life spoken and written-to-be-spoken production was still considerable.

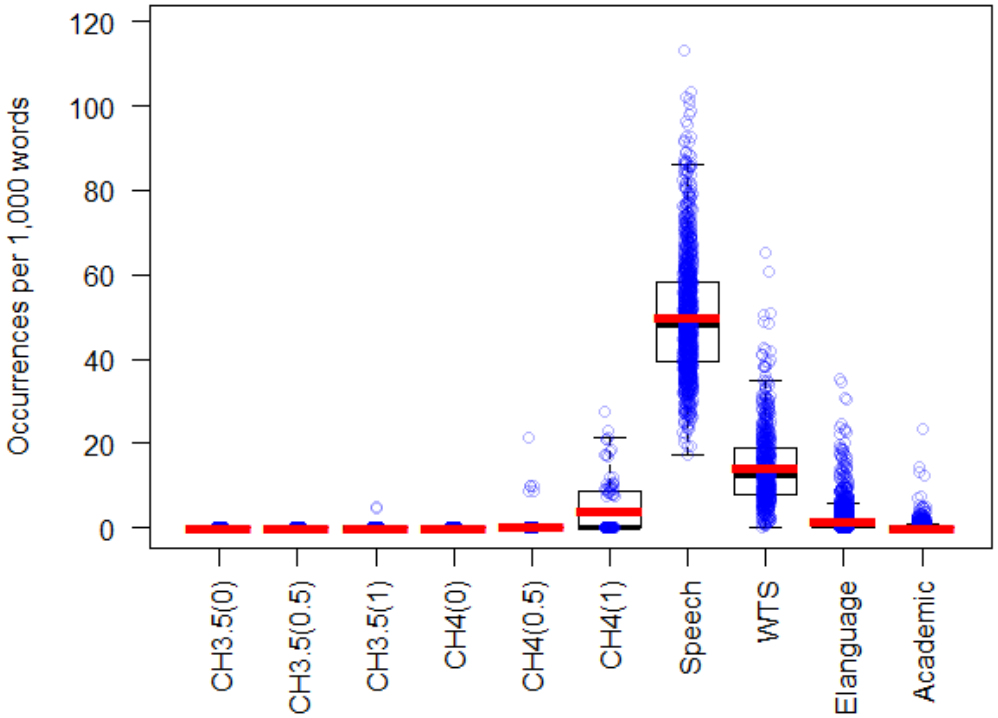

Figure 6a Pragmatic markers: Boxplots with individual data points overlaid

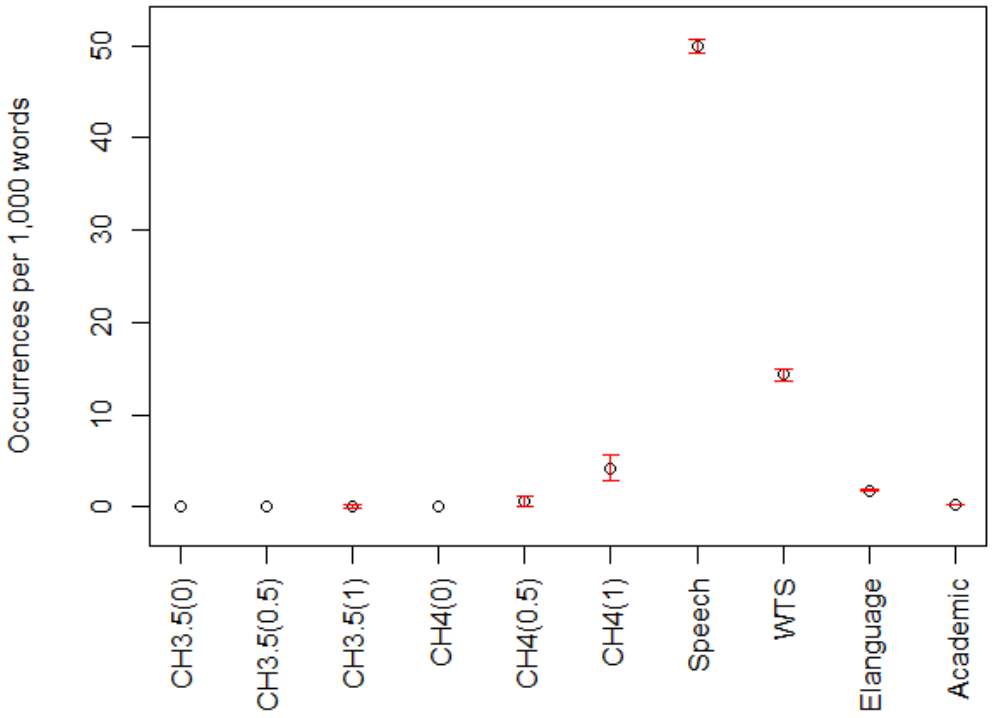

Figure 6b pragmatic markers: Means with 95% confidence intervals

Table 5 SPMs: Post-hoc Bonferroni comparisons

	CH3.5(0)	CH3.5(0.5)	CH3.5(1)	CH4(0)	CH4(0.5)	CH4(1)
CH3.5(0.5)	.001**					
CH3.5(1)	1.00	1.00				
CH4(0)	†	†	1.00			
CH4(0.5)	.915	.915	1.00	.915		
CH4(1)	.001**	.001**	.001**	.001**	.001**	
ACAD	.001**	.001**	.332	.001**	1.00	.001**
ELAN	.001**	.001**	.001**	.001**	.001**	.022*
WTS	.001**	.001**	.001**	.001**	.001**	.001**
Speech	.001**	.001**	.001**	.001**	.001**	.001**

† In the cases where there were no occurrences of the target forms observed for two chatbots, the Bonferroni corrected t-tests could not be computed.

Amplifiers and downtowners

A Welch's ANOVA showed a statistically significant effect of the text type/chatbot on amplifier and downtoner frequency [$F(11.0, 1112.9)=837.9$, $p<.001$, $\omega 2 =0.023$]; post-hoc Bonferroni comparisons are provided in Table 6. Amplifiers and downtowners are a relatively common linguistic feature across spoken and written genres, as demonstrated by the BNC2014 data (Figures 7a-b). They occur with highest frequency in elanguage (M=5.60 in 1,000 words, SD=1.67) followed by interactive speech (M=3.86, SD=1.61) and written-to-be-spoken texts (M=3.37, SD=1.67), with the lowest occurrence in academic writing (M=3.01, SD=1.52). They occurred with very low frequency in chatbot production, with only CH3.5(1) showing a somewhat higher occurrence (M=1.16, SD=2.48) with a statistically significant difference from the remaining five bots that did not differ from each other. When produced, amplifiers typically occurred in an initial position (e.g. "Absolutely, [name]!") or as part of semi-formulaic phrases such as "That's really interesting, [name]!" and "I completely agree, [name]". All genres of authentic communication contained more amplifiers and downtowners than any of the bots and this difference was statistically significant.

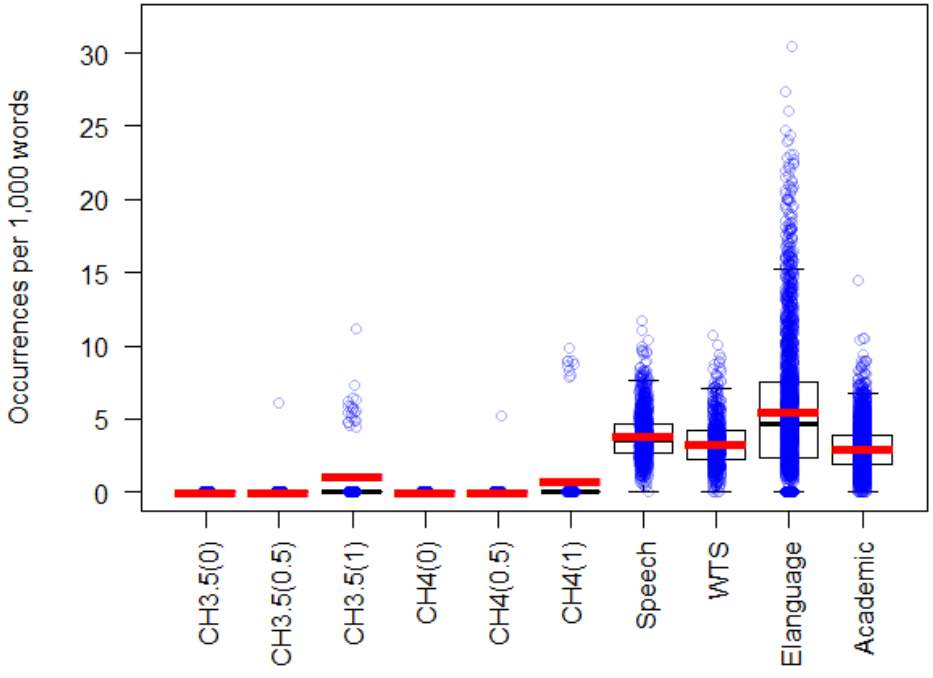

Figure 7a Amplifiers & downtoners: Boxplots with individual data points overlaid

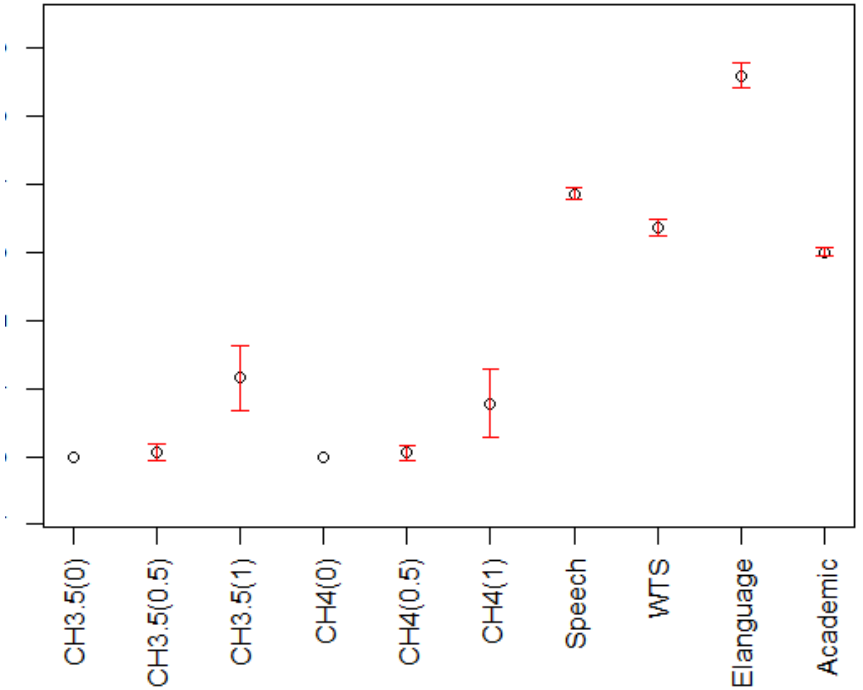

Figure 7b Amplifiers & downtoners: Means with 95% confidence intervals

Table 6 Amplifiers & downtowners: Post-hoc Bonferroni comparisons

	CH3.5(0)	CH3.5(0.5)	CH3.5(1)	CH4(0)	CH4(0.5)	CH4(1)
CH3.5(0.5)	1.00					
CH3.5(1)	.001**	.001**				
CH4(0)	†	1.00	.001**			
CH4(0.5)	1.00	1.00	.001**	1.00		
CH4(1)	.104	.268	1.00	.104	.231	
ACAD	.001**	.001**	.001**	.001**	.001**	.001**
ELAN	.001**	.001**	.001**	.001**	.001**	.001**
WTS	.001**	.001**	.001**	.001**	.001**	.001**
Speech	.001**	.001**	.001**	.001**	.001**	.001**

Micro level: Speech act of responding to an apology

This section examines how the chatbot responded to an apology for a mistake made by the other interlocutor in the simulated interactions. Although the chatbot was not explicitly instructed on how to react to apologies it consistently produced a response when a mistake or an apology occurred, making this a useful case for exploring the extent to which the chatbot production elicited in these simulations reflected patterns typical of authentic speech. Other micro-level features which are commonly found in interactive communication – such as responses to criticism, expressions of humour, opinion, empathy or solidarity – could also be investigated using the framework.

To conduct this micro-analysis, a comprehensive list of forms used by chatbots to respond to an apology or acknowledgment of a mistake was first identified through a manual analysis (see Table 7). These forms could be systematically identified because mistakes and apologies were explicitly marked (e.g., “my mistake”, “actually, I made an error”) in the other interlocutor’s production and in each case the chatbot responded to this language function. Thirty percent of the chatbot output was independently double coded by another researcher to

determine if it constituted a response to an apology/mistake, with full agreement, indicating a high level of reliability of the coding.

Overall, 281 expressions were identified in the Chatbot Corpus (see Supplementary Appendices for a full breakdown by chatbot). Some expressions appeared repeatedly (e.g. *no worries, no problem,* and *we all make mistakes*), while others occurred only once. Table 7 shows whether each form was attested at least once in the informal speech or other BNC2014 genres.

Table 7 Response to apology: Overview of forms found in chatbot production and BNC2014 subcorpora

	Chatbot Corpus	Attested in informal speech?	Attested in written genres of BNC2014?
Ah, that's quite a spin	1		
Don't worry	1	Y	WTS,F,E,N,M
[...] easy to muddle things	1		
Happens to the best of us	7		WTS,F,E,M
[...] it happens	5	Y	WTS,F,E,N,M
[...] it happens to everyone	1	Y	F
[...] it happens to us all	1		F
It's good we caught that now	1		
Mistakes happen	4		F,E,N,M
Misunderstandings happen to us all	1		
No problem	26	Y	WTS,F,E,N,M
No sweat	1	Y	WTS,F,E,N,M
No worries	193	Y	WTS,F,E,N,M
Not a problem	1	Y	WTS,F,E,N,M
Not to worry	2	Y	WTS,F,E,N,M
Oh, a twist	1		
Oh, okay	1	Y	WTS,F,E,M
Oh, that's okay	1	Y	E
Stuff happens	1	Y	F,E,N,M
Take your time	1	Y	WTS,F,E,N,M
Thanks for the clarification	4		WTS
That's a good correction	1		
That's a needed clarification	1		
That's alright	8	Y	WTS,F,E,N,M
[...] these things happen	3	Y	WTS,F,E,N,M
We all make mistakes	12		WTS,F,E,N

[WTS = Written-to-be-spoken; F = Fiction; E = Elanguage; N = Newspapers; M = Magazines]

Out of the 27 unique forms, 14 were attested in the informal speech genre of BNC2014, suggesting a good match with a real-world domain of spoken interaction. Notably, five other expressions do not occur in speech but were found in other genres – most often fiction and written-to-spoken texts. Ten expressions were not found in the BNC2014 at all (e.g. *Oh, a twist, That's a needed clarification* and *Ah, that's quite a spin*), suggesting a low level of similarity with authentic spoken interaction in some of the responses observed in the simulations.

Discussion

This study presented a novel corpus-based methodological framework for evaluating the authenticity of chatbot production. The results revealed that i) linguistic features did vary considerably across GPT version or temperature (RQ1) and ii) chatbot output in the simulations generated at this stage of development tended to differ from spoken production and to align more closely with written texts (RQ2).

Regarding the variability in chatbot production (RQ1), both MDA and comparative frequency analysis showed that occurrence of target linguistic features did not vary systematically with GPT version or temperature. For example, while the *involved-informational* MDA dimension showed a degree of temperature-related clustering across GPT versions, this pattern was not observed in other dimensions. In another example, lexical density increased at a similar rate from CH3.5(0.5) to CH3.5(1) and from CH4(0.5) to CH4(1), suggesting a linear relationship with temperature. However, chatbots with temperature 0 behaved in a less predictable manner, showing both the lowest and the highest lexical density values. These findings indicate that combining the same prompt with different system

configurations can result in largely unpredictable variation in linguistic behaviour, highlighting the importance of testing chatbot output empirically, particularly with respect to key linguistic functions and features.

Looking at the comparison of chatbot language to authentic spoken production (RQ2), across all three levels of analysis, the chatbots tended to align more closely with written genres than with spoken production in the target areas. At the macro-level, the chatbot output appeared less involved and more descriptive than informal speech, yet it also differed from more formal written genres. At the meso-level, lexical density in chatbot production was found to be similar to or higher than lexical density in two written genres. This pattern is likely related to the limited occurrence of typical spoken features such as pragmatic and stance markers revealed in some of the analyses. At the micro-level, analysis of chatbot responses to an apology showed a combination of highly conventional expressions (e.g. *no worries*), forms more typical of written genres, and some less usual responses (*oh, a twist!*) not attested in the reference corpus. The fact that some responses occurred with low frequency or not at all in (spoken) BNC2014 does not in itself mean that they are not appropriate in chatbot communication; human judgement will always be important in interpreting results at the micro-level of analysis. However, these findings suggest that this pattern (and these expressions) may require further attention and evaluation from the developers in relation to the speaking construct and the persona the chatbot is intended to reflect (given that the reference corpus selected in this study is substantial and well-matched with the parameters of the chatbot's intended persona).

The limited occurrence of some spoken features in chatbot production is likely related to the fact that generative AI systems are typically designed for, and most effective at, conveying information rather than engaging in interpersonal interaction (OECD, 2025; Dippold, 2025). Thus, even with detailed instructions regarding the nature of production (e.g., as part of the reinforcement learning from human feedback), chatbots may generate output

more typical of written or written-to-be-spoken production, reflecting the primary data types in LLM training (Dam et al., 2024). The instruction to limit typical turns to thirty-five words could have contributed to this trend, but the variation among chatbots suggests that the system configuration had a more significant impact on the prevalence of spoken features. For example, CH4(1) produced consistently more pragmatic markers than the other GPT4 versions as illustrated in Examples 1-2 in which the chatbots realised the same language function of moving the conversation to the next stage:

(1) CH4(1) Sounds like a plan, [name]. So, what's the next step?

(2) CH4(0.5) Well done! What are your next steps in producing the article?

Similarly, CH4(1) was more likely to employ spoken features in contexts such as signalling receipt of information from the other interlocutors (illustrated in Examples 3 and 4) whereas other bots tended to omit these, focusing predominately on propositional content (Example 5).

(3) CH4(1): Oh, okay. In that case, could you suggest a different job [...]

(4) CH4(1): Certainly, empathy can't be automated.

(5) CH4(0) That makes sense. Lastly, why do you think being an Architect is secure in the AI age? What's your evidence for this?

Chatbot-specific variation in the frequency of spoken features further underscores the considerable variability in chatbot linguistic behaviour and that this variability is difficult to predict based on the model attributes and system settings alone. For example, as illustrated above, the higher temperature setting in CH4 – expected to lead to more ‘creative’, ‘diverse’ or ‘random’ outputs – showed more consistency in producing some types of spoken linguistic

features than GPT-4 versions with lower temperature values. By contrast, looking at a different set of speaking features (e.g. the first MDA dimension), lower temperatures indeed seemed to be more consistent in producing the expected communicative behaviour.

This variation illustrates the opacity of reasoning within LLMs and AI-based interactive systems, posing a major challenge for their application in (high-stakes) assessment and educational contexts. Issues with opacity have been widely acknowledged, with NLP solutions being sought through, for example, increasing LLM internal transparency (e.g., Andrada et al., 2023; Suzuki et al., 2025). While such initiatives are welcome, uncertainty remains regarding their timeframe, adoption by major AI developers, and effectiveness. In the meantime, corpus methods can contribute substantially to the transparency of AI-powered chatbots, providing immediate and effective support during development phases and an evidence-based evaluation of the final products in terms of their alignment with target production. If the transparency of GenAI tools improves over time, corpus methods can play a key role in the ongoing validation of their use.

Implications and further directions

This study proposed and illustrated a systematic, multi-level framework for a corpus-based evaluation of authentic language use by chatbots designed for language education and assessment. We demonstrated that such analyses of GenAI output can be highly informative for understanding the scope and functionality of chatbots. Corpus methods provide a robust, quantitative approach to the analysis of the nature of chatbot language while also offering substantial flexibility in the aims, focus and granularity of analysis. Corpus evidence thus provides an empirical foundation to inform chatbot development and evaluation which can complement human judgements and more fine-grained analyses (e.g., through CA) of authenticity and appropriacy. Specifically, the findings of the analyses presented above could

feed into prompt refinement and further training to emphasise the importance of spoken genres. While chatbot training often relies on human judgement to evaluate outputs at different stages of the development, such feedback has typically prioritised the accuracy and coherence of chatbot turns or the ability to complete a task, with less consideration given to appropriateness in terms of the intended genre, mode or style of communication (Dippold et al., 2020). The reliability of human feedback is also often unclear given the limited information about the linguistic expertise, experience and training of human raters. The corpus-based approach helps to address these limitations.

The findings raise implications for future research directions. The differences between authentic speech and chatbot production identified during the chatbot training in this study highlight the need to broaden the research agenda on chatbot development and use in language education and assessment. Much research on AI use in these settings has focused on topics such as quality, reliability and fairness in automatic feedback provision on L2 production and chatbot interaction from the end-user perspective (e.g., the communicative behaviour and perception of learners when interacting with chatbots) (e.g., Ockey et al., 2023). However, the current study joins a growing collection of research articles that have explicitly examined the language used by *chatbots themselves* whether in educational settings or elsewhere (Dippold, 2025; Dippold et al., 2020; Eguchi et al., 2025; Voss & Waring, 2025). We therefore call for a more systematic, linguistically-grounded research agenda focused on the communicative behaviour by chatbots designed to interact with L2 learners or test-takers, including a comparison to authentic human communication.

The findings suggest deeper construct implications, as well, for the use of chatbots in spoken assessment contexts. As discussed earlier, features associated with spoken production play a crucial role in online, real-time processing and production of speech, and in managing of social and interpersonal relationships between speakers (Aijmer, 2003; Hughes & Reed,

2016; Kärkkäinen, 2003). Chatbot outputs – often characterised by more densely-packaged information and lower occurrence of discourse markers – can make it more difficult for learners and test-takers to process language during the interaction with AI-powered systems by increasing the cognitive load of the task. Moreover, such chatbot production could (systematically) expose learners and test takers to an unrealistic representation of spoken language (e.g. devoid of filled pauses, pragmatic markers, and other common spoken features) (Voss & Waring, 2025). As a result, developers and users of chatbot conversational agents in language assessment or education need to explicitly reflect on the speaking construct – guided by an understanding of key features of the target language use domain – represented by the chatbot language use. A domain definition should guide the development of chatbot systems and provide a reference point for evaluating chatbot performance. Further exploration of the speaking constructs represented by AI-powered chatbots/conversational agents is also significant from the theoretical perspective (Nakatsuhara et al., 2021), with the need to account for the increasingly complex continuum of human-human and human-machine interaction involved in language teaching and assessment tasks (Gablasova et al., 2024; Harding, 2025, Thirakunovit et al., 2019).

While this study used a large general corpus of British English (BNC2014) as a reference point, other chatbot development projects can create purpose-built corpora based on domain-relevant speaking constructs for the specific SDS/chatbots. Such specialised corpora have already been used effectively in language testing to develop and validate resources (e.g., Kyle et al., 2021; Biber et al., 2004). For example, to evaluate the appropriateness of a chatbot used as an interlocutor in an OPI-style speaking test, a corpus of OPI (human) examiner productions would be a useful reference corpus. Chatbots used for assessment in language for specific purposes (LSP) contexts such as healthcare communication or business communication would draw on relevant corpora from those domains. In this sense, the

methodological framework we have proposed is flexible and, moreover, invites developers to reflect on how authenticity can be understood in terms of domain-relevant communication in selecting suitable reference corpora.

When adapting the corpus framework proposed in this study to different contexts, the starting point for the analysis – and for the selection of appropriate linguistic features – should draw on theoretical insights and contextual considerations. We recommend beginning the process with theoretically grounded resources such as Biber's (1988) list of 67 lexico-grammatical features, which reflect distinctions across genres/registers and modalities. These features can be adapted - expanded or refined - based on the specific aims of the chatbot under development. For example, the current study placed particular emphasis on spoken features, which were used to expand the original list (Biber, 1988). The comprehensive macro-level analysis identifies feature sets sensitive to the genre or mode differences in the target and reference corpora. These findings can then guide finer-grained meso- and micro-level analyses, which should be further refined with context-specific adaptations, reflecting development and end-user needs.

The scope of this study allowed only for the inclusion of a selected set of corpus analyses to demonstrate the flexibility of the method and its possible applications. However, the corpus-based framework for chatbot output evaluation allows for a much broader range of analyses than could be shown here. For example, further corpus analyses can examine how chatbot production aligns with the intended social characteristics of the bot's personality by comparing its language with that of relevant social groups in corpora such as BNC2014 (e.g. in terms of gender, age and social class). On this point, a BNC2014 search showed that while some of the vocabulary (e.g., *definitely)* used by the chatbots was indeed typical of the younger age group (i.e., 20-29 age band) the bot was instructed to represent, other recurring words were more typical of older speakers (e.g., *certainly*). In addition, while the present study focused on

the frequency information of target linguistic features, corpus methods can further examine the diversity and distribution of chatbot language when compared to human production. As an example, the micro-level analysis of responses to an apology or a mistake could be extended further to consider their communicative appropriateness and the politeness strategies employed in them.

Conclusion

Drawing on a systematic comparison with a corpus of authentic human communication, this study examined linguistic patterns in simulated chatbot production at an early stage of the development of a low-stakes formative assessment system. It demonstrated the value of empirical language analysis of chatbots designed for language learning, teaching and assessment purposes, showing the contribution of corpus analysis to the transparency, explicability and authenticity of chatbot output. The study also demonstrates the potential of integrating corpus analyses into the chatbot development cycle, providing evidence-based guidance for SDS development. The methodological framework presented in this paper is applicable to other chatbot models and remains relevant even as AI tools continue to evolve. Evaluating chatbot authenticity, as we have argued above, requires a multi-layered comparison with the linguistic characteristics of a representative dataset drawn from the target language use domain. The corpus-based approach is uniquely positioned to allow for detailed and efficient comparative work. The framework proposed in this study therefore represents a crucial foundation for establishing standard and rigorous procedures for evaluating and reporting the quality and functionalities of AI systems used in language education and assessment.

References

Aijmer, K. (2013). *Understanding pragmatic markers: A variational pragmatic approach*. Edinburgh University Press. https://doi.org/10.1515/9780748635511

Andrada, G., Clowes, R. W., & Smart, P. R. (2023). Varieties of transparency: Exploring agency within AI systems. *AI & society*, *38*(4), 1321-1331. https://doi.org/10.1007/s00146-021-01326-6

Bachman, L. F. (1991). What does language testing have to offer? *TESOL Quarterly*, 25(4), 671-704. https://doi.org/10.2307/3587082

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, *2*(1), 1-34. https://doi.org/10.1207/s15434311laq0201_1

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press. https://doi.org/10.1017/cbo9780511621024

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E. and Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus*. TOEFL Monograph Series. Educational Testing Service.

Biber, D., & Conrad, S. (2019). *Register, genre, and style*. Cambridge University Press. https://doi.org/10.1017/9781108686136

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press. https://doi.org/10.1017/9781316410899

Brezina, V., Hawtin, A., & McEnery, T. (2021). The Written British National Corpus 2014–design and comparability. *Text & Talk*, *41*(5-6), 595-615. https://doi.org/10.1515/text-2020-0052

Brezina, V. & Platt, W. (2024) *#LancsBox X* [software], Lancaster University, http://lancsbox.lancs.ac.uk.

Caines, A., McCarthy, M., & O'Keeffe, A. (2016). Spoken language corpora and pedagogical applications. In F. Farr & L. Murray (Eds.) *The Routledge handbook of language learning and technology* (pp. 348-361). Routledge. https://doi.org/10.4324/9781315657899

Chalhoub-Deville, M., & O'Sullivan, B. (2020). *Validity: Theoretical development and integrated arguments*. University of Toronto Press. https://doi.org/10.3138/9781781799918

Chapelle, C. A. (2025). Generative AI as game changer: Implications for language education. *System*, *132*, 103672. https://doi.org/10.1016/j.system.2025.103672

Council of Europe (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion Volume*. Language Policy Programme Education Policy Division Education Department, Council of Europe. https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

Dam, S. K., Hong, C. S., Qiao, Y., & Zhang, C. (2024). A complete survey on LLM-based AI chatbots. Computers and Language. *arXiv preprint arXiv:2406.16937*.

Dippold, D. (2025). Making the case for audience design in conversational AI: Users' pragmatic strategies and rapport expectations in interaction with a task-oriented chatbot. *Applied Linguistics*, *46*(4), 650-667. https://doi.org/10.1093/applin/amae033

Dippold, D., Mold, F., & Ghosh, P. (2025). Improving chatbot design and intent recognition: An approach through the methods of intercultural pragmatics. *Intercultural Pragmatics*, *22*(2), 439-458. https://doi.org/10.1515/ip-2025-2010

Dippold, D., Lynden, J., Shrubsall, R., & Ingram, R. (2020). A turn to language: How interactional sociolinguistics informs the redesign of prompt: response chatbot turns. *Discourse, Context & Media*, *37*, 100432. https://doi.org/10.1016/j.dcm.2020.100432

Davis, J., Van Bulck, L., Durieux, B. N., & Lindvall, C. (2024). The Temperature Feature of ChatGPT: Modifying Creativity for Clinical Research. *JMIR Human Factors, 11*, e53559. https://doi.org/10.2196/53559

Egbert, J., Biber, D., & Gray, B. (2022). *Designing and evaluating language corpora: A practical framework for corpus representativeness*. Cambridge University Press. https://doi.org/10.1017/9781316584880

Eguchi, M., Takizawa, K., Saeki, M., Kurata, F., Suzuki, S., Matsuyama, Y., & Sawaki, Y. (2025). Human-versus artificial intelligence-delivered roleplay tasks for assessing interactional competence: An applied conversation analytic study. *TESOL Quarterly*, 59, 183-219. https://doi.org/10.1002/tesq.70028

Gablasova, D. (2021). Corpora for second language assessments. In Brunfaut, T. & Winke, P. (Eds). *The Routledge handbook of second language acquisition and language testing* (pp. 45-53). Routledge. https://doi.org/10.4324/9781351034784-6

Gablasova, D., Harding, L., Brezina, V., & Dunlea, J. (2024). Expressions of epistemic stance in computer-mediated L2 speaking assessment: A corpus-based approach. *International Journal of Learner Corpus Research*, *10*(1), 183-215. https://doi.org/10.1075/ijlcr.00044.gab

Harding, L. (2025). Utopian and dystopian visions: Steering a course for the responsible use of artificial intelligence (AI) in language testing and assessment. *Language Testing*, *42*(4), 561-575. https://doi.org/10.1177/02655322251350717

Hasrol, S. B., Zakaria, A., & Aryadoust, V. (2022). A systematic review of authenticity in second language assessment. *Research Methods in Applied Linguistics*, *1*(3), 100023. https://doi.org/10.1016/j.rmal.2022.100023

Hazelhurst, E.T., Gablasova, D. & Wilson, M. (2024). Chatbot Corpus. *Electronic dataset.*

Hughes, R., & Reed, B. S. (2016). *Teaching and researching speaking*. Routledge. https://doi.org/10.4324/9781315692395

Jablonkai, R. R., & Csomay, E. (Eds.). (2022). *The Routledge handbook of corpora and English language teaching and learning*. Taylor & Francis. https://doi.org/10.4324/9781003002901

Jeon, J., Lee, S., & Choe, H. (2023). Beyond ChatGPT: A conceptual framework and systematic review of speech-recognition chatbots for language learning. *Computers & Education*, 206, 104898. https://doi.org/10.1016/j.compedu.2023.104898

Karatay, Y. & Xu, J. (2025). Exploring the potential of conversational AI for assessing second language oral proficiency. *TESOL Quarterly*, 59, 220-250. https://doi.org/10.1002/tesq.70003

Kärkkäinen, E. (2003). *Epistemic stance in English conversation*. John Benjamins. https://doi.org/10.1075/pbns.115

Knoth, N., A. Tolzin, & A. Janson (2024). AI literacy and its implications for prompt engineering strategies. Computers and Education: Artificial Intelligence, 49(4), 100225. https://doi.org/10.1016/j.caeai.2024.100225

Krook, J., Winter, P., Downer, J., & Blockx, J. (2025). A systematic literature review of artificial intelligence (AI) transparency laws in the European Union (EU) and United Kingdom (UK): a socio-legal approach to AI transparency governance. *AI and Ethics*, 1-22. https://doi.org/10.1007/s43681-025-00674-z

Kyle, K., Choe, A., Eguchi, M., LaFlair, G. & Ziegler, N. (2021). A comparison of spoken and written language use in traditional and technology-mediated learning environments. *ETS Research Report Series* 2021. https://doi.org/10.1002/ets2.12329

Lee, M. (2023). A Mathematical Investigation of Hallucination and Creativity in GPT Models. Mathematics, 11(10), 2320. https://doi.org/10.3390/math11102320

Li, L., Sleem, L., Gentile, N., Nichil, G., & State, R. (2025). Exploring the Impact of Temperature on Large Language Models: Hot or Cold? Procedia Computer Science, 264, 242–251. https://doi.org/10.1016/j.procs.2025.07.135

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, *22*(3), 319-344. https://doi.org/10.1075/ijcl.22.3.02lov

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press. https://doi.org/10.1017/cbo9780511981395

McEnery, T., & Brezina, V. (2022). *Fundamental principles of corpus linguistics*. Cambridge University Press. https://doi.org/10.1017/9781107110625

Moorhouse, B. L., & Wong, K. M. (2025). *Generative Artificial Intelligence and Language Teaching*. Cambridge Elements. Cambridge University Press. https://doi.org/10.1017/9781009618823

Nakatsuhara, F., Khabbazbashi, N., & Inoue, C. (2021). Assessing speaking. In Fulcher, G. & Harding, L. (Eds.) *The Routledge handbook of language testing* (pp. 209-222). Routledge. https://doi.org/10.4324/9781003220756-17

Ockey, G. J., Chukharev-Hudilainen, E., & Hirch, R. R. (2023). Assessing interactional competence: ICE versus a human partner. *Language Assessment Quarterly*, *20*(4-5), 377-398. https://doi.org/10.1080/15434303.2023.2237486

OECD. (2025). *Introducing the OECD AI Capability Indicators.* OECD Publishing. https://doi.org/10.1787/be745f04-en

OpenAI. (2024a). *OpenAI API Reference*. Retrieved March 19, 2024, from https://platform.openai.com/docs/api-reference/chat/create

OpenAI. (2024b). *Introducing Chat GPT*. Retrieved March 19, 2024, from https://openai.com/index/chatgpt/

OpenAI. (2024c). *GPT-4*. Retrieved March 19, 2024, from https://openai.com/research/gpt-4

O'Sullivan, B. (2021). *The Comprehensive Learning System*. British Council. https://doi.org/10.57884/ZDEC-CK78

Peeperkorn, M., Kouwenhoven, T., Brown, D., & Jordanous, A. (2024). Is temperature the creativity parameter of large language models? In K. Grace, M. T. Llano, P. Martins, & M. M. Hedblom (Eds.), Proceedings of the 15th International Conference on Computational Creativity, ICCC 2024, Jönköping, Sweden, June 17-21, 2024 (pp. 226–235). https://doi.org/10.48550/arXiv.2405.00492

R Core Team (2025). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Radziwill, N. M., & Benton, M. C. (2017). Evaluating quality of chatbots and intelligent conversational agents. *arXiv preprint arXiv:1704.04579*.

Renze, M., & Guven, E. (2024). The effect of sampling temperature on problem solving in large language models. *Proceedings of Findings of the Association for Computational Linguistics: EMNLP,* pp. 7346–7356, https://doi.org/10.18653/v1/2024.findings-emnlp.432

Roever, C., & Ikeda, N. (2022). What scores from monologic speaking tests can (not) tell us about interactional competence. *Language Testing*, *39*(1), 7-29. https://doi.org/10.1177/02655322211003332

Saeki, M., Takatsu, H., Kurata, F., Suzuki, S., Eguchi, M., Matsuura, R., Takizawa, K., Yoshikawa, S., & Matsuyama, Y. (2024). InteLLA: Intelligent Language Learning Assistant for Assessing Language Proficiency through Interviews and Roleplays. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 385399). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.sigdial-1.34

Timpe-Laughlin, V., Sydorenko, T., & Dombi, J. (2024). Human versus machine: investigating L2 learner output in face-to-face versus fully automated role-plays. *Computer Assisted Language Learning*, *37*(1–2), 149–178. https://doi.org/10.1080/09588221.2022.2032184

Thirakunkovit, S., Rodríguez-Fuentes, R. A., Park, K., & Staples, S. (2019). A corpus-based analysis of grammatical complexity as a measure of international teaching assistants' oral English proficiency. *English for Specific Purposes*, *53*, 74-89. https://doi.org/10.1016/j.esp.2018.09.002

Voss, E., & Waring, H. Z. (2025). When ChatGPT can't chat: The quest for naturalness. *TESOL Quarterly*, *59*(2), 1064-1075. https://doi.org/10.1002/tesq.3374

Wagner, E. (2014). Using unscripted spoken texts in the teaching of second language listening. *TESOL journal*, *5*(2), 288-311. https://doi.org/10.1002/tesj.120

Xi, X. (2023). Advancing language assessment with AI and ML–leaning into AI is inevitable, but can theory keep up? *Language Assessment Quarterly*, 20(4–5), 357–376. https://doi.org/10.1080/15434303.2023.2291488