RESEARCH ARTICLE

How Authentic is AI? Comparing AI and Human-Authored EFL Listening Materials

Received: 05 November 2025; Revised: 10 March 2026; Published: 14 May 2026

Mark Donnellan

Kindai University, Osaka, Japan
Email: donnellan@kindai.ac.jp
ORCID: https://orcid.org/0009-0005-5318-6683

Abstract:

This study investigates the linguistic authenticity of AI-generated English as a Foreign Language (EFL) listening materials through a corpus analysis comparing texts produced by ChatGPT-5, Gemini 2.5, and Claude 4.5 against human-authored materials. Building on a pilot study of ChatGPT-4, this research examines how various LLMs replicate spoken discourse features essential for EFL listening and pragmatic skills. Using Sketch Engine, the study analyzed four corpora totaling approximately 101,000 tokens, examining lexical variety, n-gram patterns, and discourse marker usage. Results reveal that while AI-generated texts demonstrate higher type-token ratios (0.128-0.130 vs. 0.107), they significantly underrepresent conversational features crucial for authentic interaction. Human-authored materials contained 50% discourse markers among top keywords compared to 10-25% in AI outputs ($\chi^2 = 8.97$, $p = 0.030$). Analysis showed AI corpora, particularly ChatGPT and Claude, exhibited 50-61% formulaic language typical of scripted presentations, contrasting with conversational variability in human texts. Discourse marker frequency was significantly higher in human materials (2.52%) than AI-generated texts (1.15-1.45%). These findings suggest that current LLMs produce language resembling scripted discourse rather than authentic dialogue, limiting their effectiveness for developing listening skills. The study concludes that AI-generated materials require careful supplementation with authentic materials to meet learners' communicative needs. Implications emphasize balanced integration of AI tools with human-authored content and corpus-based evaluation methods for assessing AI-generated educational materials.

Keywords: AI-generated text, corpus linguistics, linguistic authenticity, EFL listening materials, large language models, discourse markers, pragmatic competence

1. Introduction

The emergence of AI-trained large language models (LLMs), developed by leading AI development organizations, has fueled both intense scrutiny and great enthusiasm in educational and academic circles. Prominent models such as GPT-5 (OpenAI, 2025), Claude 4.5 (Anthropic, 2025), Gemini 2.5 (Google DeepMind, 2025), Llama 4 (Meta, 2025), and Grok-3 (xAI, 2025), have demonstrated a varied range of strengths. These LLMs generate naturalistic language and support academic and pedagogical contexts, as demonstrated in a pilot study by the current researcher (Donnellan, 2025). Since 2023, rapid advancements have transformed their capabilities. These improved capabilities include Grok-3's real-time reasoning and multimodal integration via X, which enhance dynamic dialogue generation (xAI, 2025), Claude 4.5's task-specific tools, such as skills and life sciences integration, prioritize safe

and structured outputs (Anthropic, 2025), Llama 4's 256K context length and native multimodality enable complex text processing (Meta, 2025), and Gemini 2.5's audio-visual processing supports rich EFL material creation (Google DeepMind, 2025). These developments amplify LLMs' potential for English as a Foreign Language (EFL) pedagogy yet raise critical questions about their linguistic authenticity and suitability for classroom use (Liu et al., 2024).

One particularly promising yet contentious application is the development of pedagogical materials for EFL learners. LLMs offer scalable solutions for generating dialogues, exercises, and feedback, but their outputs must align with the communicative needs of learners, particularly in developing listening skills and pragmatic competence (Wei, 2023). The pilot study (Donnellan, 2025) employed corpus-linguistic methods to compare human-authored EFL listening dialogues with outputs generated by ChatGPT-4, utilizing the subscription-based version available during corpus compilation in 2023. That study revealed that ChatGPT-4 produced grammatically accurate texts with considerable lexical variety but exhibited notable deficiencies in conversational authenticity, including spontaneity and discourse markers like hesitations and fillers. These findings highlighted the limitations of relying solely on AI-generated texts for naturalistic language exposure, a critical component of EFL instruction.

Given the rapid evolution of LLMs and their diverse architectures, single-model studies like the pilot study for this project are increasingly insufficient for capturing the variability in AI-generated language (Seo et al., 2025). Post-2023 advancements—such as Grok-3's real-time knowledge integration, Claude 4.5's ethical output enhancements, and Gemini 2.5's multimodal capabilities—demand a broader evaluation to assess their suitability for EFL contexts (Zheng et al., 2025). While recent reviews highlight LLMs' educational potential, few studies empirically compare multiple models against human-authored materials in EFL settings, particularly for listening tasks (Goh & Aryadoust, 2025). This study extends the 2023 corpus analysis to free versions of ChatGPT-5, Gemini 2.5, and Claude 4.5—tools accessible to teachers. These models were selected for their varied strengths: ChatGPT-5's ubiquity, and Claude 4.5's safety-focused outputs (OpenAI, 2025; Google DeepMind, 2025; Anthropic, 2025). By using free tiers, this study mirrors real-world pedagogical practices, aiming to generate texts that are similar to those which EFL materials developers may currently be using LLMs to produce.

This research is guided by the following questions:

> RQ1 What linguistic differences in lexical variety, n-grams, and discourse markers exist between texts generated by ChatGPT-5, Gemini 2.5, Claude 4.5, and human-authored EFL listening materials?
>
> RQ2 What are the implications of these differences for EFL materials development?

Building on prior findings (Donnellan, 2025), these questions explore multi-model variability to inform EFL material design, extending the pilot study's framework with actionable insights for integrating diverse LLMs into language education.

2. Literature Review

The rapid advancement of LLMs has transformed educational research and practice, particularly in EFL pedagogy. This literature review synthesizes current scholarship on LLMs, focusing on their development, applications in EFL, linguistic characteristics, comparative analyses, and corpus

linguistics methods. It builds on prior work (Donnellan, 2025) to contextualize a multi-model corpus analysis comparing ChatGPT-5, Gemini 2.5, Claude 4.5, and human-authored EFL listening materials, focusing mainly on their suitability for developing conversational listening skills and pragmatic competence to assist the development of communicational skills.

2.1 Evolution and diversity of LLMs

Since their mainstream emergence, LLMs such as ChatGPT (OpenAI, 2025), Claude (Anthropic, 2025), Gemini (Google DeepMind, 2025), and Llama (Meta, 2025) have evolved significantly, with 2024–2025 iterations introducing advanced capabilities. ChatGPT-5 offers enhanced reasoning and multimodality, enabling seamless integration of text, images, and voice for dynamic interactions (OpenAI, 2025). Claude 4.5 builds on prior versions with superior agentic performance, including improved tool handling, memory management, and multimodal reasoning for tasks like coding and complex analysis (Anthropic, 2025). Gemini 2.5, infused with LearnLM, excels in educational personalization through its 1-million-token context window and native multimodality, supporting audio-visual processing for immersive learning (Google DeepMind, 2025), while Llama 4's 256K context length and native multimodality support complex text processing (Meta, 2025). These advancements, driven by diverse architectures like mixture-of-experts and expanded training datasets, underscore the need for multi-model evaluations to capture variability in outputs (Hajikhani, 2024). Recent reviews highlight how advanced multimodal language models outperform earlier versions in personalization and reasoning capabilities, making them increasingly promising for educational contexts, particularly where adaptive learning and multimodal input processing can enhance learner engagement (Bewersdorff et al., 2025; Kasneci et al., 2023). For instance, multimodal AI systems' ability to process and generate integrated audio-visual content holds potential for enhancing vocabulary acquisition and pronunciation practice in EFL contexts (Huang et al., 2023). However, rapid updates render single-model studies, such as early ChatGPT analyses, increasingly outdated.

2.2 Applications and opportunities in EFL pedagogy

LLMs have reshaped EFL pedagogy by enabling scalable content generation, personalized feedback, and interactive learning. Kasneci et al. (2023) note their potential for creating dialogues, exercises, and automated feedback, reducing teacher workload while enhancing accessibility. Recent developments in multimodal AI systems have expanded these capabilities, with advanced models facilitating tasks such as lesson planning, content adaptation, and differentiated instruction across proficiency levels (Bewersdorff et al., 2025). The integration of AI tools into educational platforms, particularly through learning management systems and collaborative workspace environments, has increased accessibility for educators seeking to incorporate technology-enhanced materials into their teaching practice (Google Workspace, 2025). Such integration enables the creation of multimodal resources including audio examples, visual aids, and interactive exercises that can be tailored to individual learner needs.

Free-tier availability of advanced AI tools aligns with real-world teacher practices, democratizing access to technology-enhanced instruction and supporting differentiated approaches in diverse classroom contexts. However, Chen et al. (2024) caution against over-reliance, citing risks of student dependency and reduced critical thinking. A pilot study conducted by the researcher (Donnellan, 2025) found that while ChatGPT-4 demonstrated lexical variety, its lack of conversational authenticity limited its usefulness for listening tasks designed to develop learners' understanding of spoken

interaction—a concern that remains relevant as newer models continue to struggle with replicating the spontaneity and pragmatic features of natural spoken discourse (Sandler et al., 2024). These findings highlight the need for balanced integration of AI-generated and human-authored materials to meet EFL learners' communicative needs, with careful attention to the limitations of current systems in capturing authentic interactional patterns essential for developing pragmatic competence.

2.3 Linguistic characteristics

The linguistic output of LLMs is a critical focus for EFL research, as authenticity in listening materials is essential for pragmatic competence. Research has demonstrated that while AI-generated texts can exhibit lexical diversity, they often lack the spontaneity and natural variation characteristic of human speech (Sandler et al., 2024). They compared ChatGPT dialogues to human conversations, noting homogeneity and emotional neutrality in AI outputs, which may hinder learners' exposure to naturalistic discourse features essential for developing communicative competence.

The linguistic limitations of AI-generated content extend to several dimensions relevant to language learning. Studies examining AI dialogue systems have found that they frequently produce outputs lacking in emotional depth and conversational authenticity, potentially limiting their effectiveness for developing pragmatic skills (Chen et al., 2024). Furthermore, AI-generated texts often exhibit reduced variability in discourse features such as hesitations, false starts, and hedging devices that characterize spontaneous spoken interaction (Kasneci et al., 2023). These discourse markers are crucial for EFL learners to recognize and comprehend, as they frequently occur in authentic listening contexts and contribute to pragmatic understanding.

Research on multimodal AI systems suggests potential improvements in generating more naturalistic language. Bewersdorff et al. (2025) note that multimodal LLMs demonstrate enhanced capabilities in processing and generating integrated audio-text outputs, which may benefit language learning applications requiring authentic listening materials. However, significant gaps remain in empirical research examining whether these systems can adequately replicate the phonetic variation, prosodic features, and interactional patterns found in natural speech. As noted in the pilot study (Donnellan, 2025), ChatGPT-4 struggled to produce the full range of discourse markers and spontaneous speech features that characterize authentic spoken English. These shortcomings have limited the ability of LLMs to replicate the unpredictability of authentic spoken English, which remains crucial for developing EFL listening comprehension skills.

A fundamental principle of spoken language is that linguistic choices are governed by communicative context and register (Biber et al., 1999; Carter & McCarthy, 2006). In informal conversation—the type of interaction represented by the dialogue subcorpus in this study—speakers engage in real-time, co-constructed discourse characterized by spontaneity, dysfluency, and interactive management (O'Keeffe et al., 2007). Features such as hesitation markers (*um, uh*), fillers (*like, you know*), backchannels (*yeah, mm-hmm*), repairs, and elliptical structures emerge naturally as participants negotiate meaning on the fly (Tao, 2003; Koester, 2010a). These elements are essential for developing pragmatic and listening competence in EFL learners, as they reflect authentic interactional norms (McCarthy & Carter, 2014).

In contrast, presentation scripts—the monologue component of the corpus—represent planned, careful speech. Speakers typically prepare in advance, resulting in higher lexical density, fewer

dysfluencies, more complex syntax, and explicit discourse structuring (e.g., *firstly, in conclusion, to summarize*) (Biber et al., 1999). This register prioritizes clarity, coherence, and informational accuracy over interactivity.

2.4 Comparative analyses of multiple LLMs

While single-model studies dominate literature, multi-model comparisons are emerging to address variability in large language model outputs and their educational applications. Recent comparative evaluations have examined multiple AI systems across various educational benchmarks. For instance, a systematic comparison of GPT-4, Claude 3 Opus, and Gemini 1.0 Ultra on undergraduate-level control engineering problems found that Claude 3 Opus emerged as the state-of-the-art model for solving complex technical problems (Schillinger et al., 2024). In medical education contexts, advanced language models demonstrated strong performance on the Japanese National Medical Licensing Examination, with GPT-4o achieving the highest accuracy rate of 89.2%, outperforming Claude 3 Opus, Gemini 1.5 Pro, and GPT-4 (Liu et al., 2025).

In educational assessments more broadly, comparative studies have shown that advanced language models significantly outperform human benchmarks in undergraduate-level knowledge and reading comprehension tasks, though performance varies across specific cognitive skills (Yilmaz et al., 2024). A zero-shot evaluation across 76 spatial tasks revealed that while chatbots performed well on spatial literacy and Geographic Information Systems theory, they exhibited weaknesses in mapping, code writing, and spatial reasoning (Hochmair et al., 2024). These findings underscore the importance of task-specific evaluations when selecting AI tools for educational contexts.

However, research specifically comparing LLMs for EFL listening materials remains limited. Earlier corpus analysis (Donnellan, 2025) examining ChatGPT-4 and human-authored texts revealed deficits in linguistic features such as n-grams and spontaneity markers characteristic of authentic speech. The paucity of multi-model studies in EFL contexts, particularly those addressing listening comprehension materials, necessitates further empirical investigation. Such research is essential for informed pedagogical decisions about AI-generated material design, especially given evidence that different models exhibit distinct strengths in areas such as ethical reasoning, personalization, and multimodal processing (Chen et al., 2024; Schillinger et al., 2024).

2.5 Corpus linguistics for evaluating AI outputs

Corpus linguistics provides a robust framework for analyzing LLM outputs, using tools like Sketch Engine to examine lexical variety, n-grams, and discourse patterns (Kilgarriff et al., 2014; McEnery & Hardie, 2012). These computational approaches enable systematic comparison of AI-generated texts against authentic human language corpora, revealing patterns that may not be immediately apparent through qualitative analysis alone (Biber & Conrad, 2019). Crosthwaite and Baisa (2023) explored integrating LLMs into corpus analysis, finding AI-assisted methods promising but prone to biases. Recent work demonstrates how corpus tools can detect distinctive features of AI-generated texts in EFL writing, enhancing rigor in linguistic evaluations and identifying characteristics that differentiate machine-generated from human-produced language (Crosthwaite & Baisa, 2023).

Corpus-based methods have proven particularly valuable for analyzing grammatical constructions and discourse features in AI outputs (Römer, 2020). Such approaches are applicable to listening texts,

where extended interactions require analysis of turn-taking patterns, discourse markers, and conversational features that characterize authentic spoken language. Donnellan (2025) employed Sketch Engine to compare ChatGPT-4 and human corpora, identifying authenticity gaps in lexical diversity and spontaneous speech features. Extending this approach to multiple LLMs requires careful consideration of methodological constraints, including tool limitations such as cost and scalability, as well as the potential for hybrid approaches that combine AI-assisted analysis with traditional corpus methods (Gries, 2021). As Leech (2000) argues, corpus-linguistic methods must balance computational efficiency with theoretical rigor to ensure findings are both statistically robust and pedagogically meaningful.

2.6 Corpus size and methodological considerations

The current study reuses the small-scale specialized human corpus employed in the pilot study. Small corpora offer practical advantages for researchers with constrained resources while maintaining analytical validity when appropriately designed. Koester (2010b) argues that carefully constructed small corpora, when representative of the target domain, yield reliable insights into linguistic patterns without requiring massive datasets. This perspective aligns with Nelson's (2010) observation that corpus size must be balanced against research objectives, with smaller, specialized corpora often proving more suitable for focused investigations than larger, general-purpose collections. For EFL research, small corpora enable focused analyses of specific genres, such as listening dialogues, which are critical for assessing pragmatic competence (McEnery & Hardie, 2012; O'Keeffe et al., 2007).

The pilot study (Donnellan, 2025), with a human corpus of 18,539 tokens, demonstrated that small-scale analyses can effectively identify differences in lexical variety and discourse markers, supporting their suitability for evaluating AI outputs. This corpus size falls within the range that Flowerdew (2004) considers appropriate for specialized corpus studies, where depth of analysis and domain specificity take precedence over breadth of coverage. Recent research has further highlighted how small corpora facilitate rapid, cost-effective studies using accessible tools like Sketch Engine, making corpus-based research feasible for educational researchers with limited computational resources (Gablasova et al., 2017).

Despite these benefits, small corpora have acknowledged limitations. McEnery and Wilson (2001) caution that their representativeness may be constrained, potentially missing broader linguistic trends observable in larger datasets. However, as Flowerdew (2004) emphasizes, careful corpus design and clear delimitation of the target domain can mitigate these concerns, ensuring relevance to specific research questions. For EFL listening tasks, small-scale corpora remain a practical and robust methodological choice, enabling fine-grained analysis of task-specific linguistic features while maintaining ecological validity (Adolphs & Knight, 2010). Such corpora are particularly suited to investigations of spoken language features, where the quality of transcription and contextual annotation often outweigh the advantages of raw size (Carter & McCarthy, 2017).

The current study capitalizes on the strengths of small-scale corpora to yield practical, context-specific insights into AI-generated texts, thereby reinforcing the value of such corpora in applied linguistics research. Building on the pilot study comparing ChatGPT-4 and human-authored materials, it addresses a key gap in the literature by directly contrasting the outputs of multiple leading LLMs—specifically, current versions of ChatGPT, Gemini, and Claude—with authentic EFL texts. This comparative approach offers evidence-based guidance for the effective integration of AI-generated

materials into language education, while providing methodological insights into corpus-based evaluation of emerging technologies in applied linguistics contexts. The following sections detail the research questions, corpus design, and analytical procedures employed to conduct this comparison, with particular attention to the lexical, grammatical, and discourse features that distinguish AI-generated from human-authored EFL listening materials.

3. Methodology

3.1 The four corpora

The current study reuses an updated version of the human corpus used in the pilot study, which consisted of 18,539 tokens. Slight alterations and additions to the corpus were made, including the addition of short speeches from two of the textbooks and the updating of dialogues, with the texts for this study taken from the latest editions of the textbooks. The texts were taken from six EFL textbooks written by the same authors. The researcher obtained written permission from the publishers to use these materials for the research project. All textbooks purported to adopt a task-based language teaching approach (Willis & Willis, 2007) and were designed with an emphasis on maintaining authentic spoken interaction. The authors recorded unscripted dialogues and did not edit or record the material to remove errors, hesitations, or other features of spontaneous speech, including those produced by non-native speakers. As a result, the dialogues preserve many features of spontaneous spoken interaction and closely resemble naturally occurring conversation. The textbooks contain speech from both native and non-native speakers of English. Three of the textbooks focused on oral communication, from which a total of 36 pair dialogues were selected. One textbook focused primarily on reading but contained listening sections, from which nine pair dialogues were taken. These dialogues from four textbooks were chosen because they contained the linguistic characteristics of informal conversation described in Section 2.3 above. The remaining two textbooks focused on presentation skills and featured monologues displaying the linguistic characteristics of presentations outlined in Section 2.3. A total of 20 longer presentation scripts (350–500 words) and 14 shorter ones (150–200 words) were taken from these books. Table 1 provides an overview of the textbooks used.

Table 1. Overview of the Textbooks

	CEFR Level	Skills Focus	Texts Used in the Corpus
Textbook A (Publisher A)	A2	oral communication	twelve pair dialogues
Textbook B (Publisher A)	B1	oral communication	twelve pair dialogues
Textbook C (Publisher A)	B2	oral communication	twelve pair dialogues
Textbook D (Publisher B)	not stated; estimated at B1	reading	nine pair dialogues
Textbook E (Publisher A)	A2	presentation	ten full presentations eight short presentations
Textbook F (Publisher A)	B1	presentation	ten full presentations six short presentations

***Publisher permission was granted for use in this study

The ChatGPT corpus used for the pilot study was not reused. This was because it was created in 2023 using ChatGPT-4. Given the rapid advancements in the field of AI, AI-generated texts from 2023 might be considered obsolete. For this study, three new corpora were created: a corpus of texts created using Claude 4.5, one created using ChatGPT-5, and one created using Gemini 2.5. The goal was to prompt these three LLMs to create texts based on the same topics as the human texts and to prompt them in such a way that they might include some of the features of natural spoken language and presentations that were not evident in the pilot study. Exact replication was avoided to enable meaningful comparison across corpora. In addition, the texts were to be in line with the CEFR levels in Table 1 above and about the same length as the texts used to create the human corpus. The first step was to create summaries of each text. In order to do this, the researcher used an LLM that was not to be used for corpus creation, Microsoft Copilot (2025). The rationale for this was to avoid any bias or training of the three LLMs to be used for the study that might occur as a result of giving them the human corpus. The researcher input all 79 texts into Microsoft Copilot and prompted it to create short summaries. These summaries were then used, and the three LLMs that are the focus of this study were asked to create texts of appropriate length (word count) and CEFR level. In addition, the LLMs were asked to include some of the features that one might expect to see in natural spoken English. As outlined in Section 2.3, these include discourse markers such as hesitations and fillers. In many cases, the LLMs produced texts that were significantly shorter than the requested word count necessitating additional prompting.

3.2 Corpus analysis

A corpus analysis was carried out on the four corpora using Sketch Engine (Kilgarriff et al., 2014). This analysis aimed to elucidate the range and variety of language that the three LLMs could produce and how the language they produced compared to the language in the human corpus which closely resembles naturally occurring conversation. The analysis focused on three aspects of the data: the number of unique words (type count), the most frequent words (keywords), and recurring multi-word sequences (n-grams). The data on the corpora that were compiled and the results of the corpus analysis can be seen in Section 4. Below is a detailed outline of the analysis:

Sketch Engine was selected due to its robust support for multi-word unit extraction, keyword analysis, and comparability across small corpora. All texts were uploaded as plain text files and processed with lemmatization disabled to preserve morphological variation typical in spoken language (e.g., *goin'* vs. *going*), and stop words were retained for n-gram analysis because function words are integral to recurring conversational sequences. The human corpus contained 25,653 tokens; the AI-generated corpora were analyzed separately: Claude 4.5 (27,375 tokens), Gemini 2.5 (26,494 tokens), and ChatGPT-5 (21,931 tokens). Clusters of three-word and four-word sequences (n-grams) were extracted from each corpus, applying a minimum frequency threshold of three occurrences to reduce noise, consistent with standard practices in spoken corpus research (e.g., Biber et al., 1999). Keywords were generated using Sketch Engine's built-in function with the EnTenTen2021 as the reference to highlight deviations from naturalistic speech.

In summary, the corpus construction and analysis were designed to enable a systematic, replicable comparison between human-authored EFL materials and AI-generated equivalents, with careful controls for text length, CEFR alignment, and prompting to elicit naturalistic spoken features. The use

of Sketch Engine and standardized analytical parameters ensures transparency and consistency across corpora. The results of this analysis are presented in Section 4.

4. Results

4.1 Corpus overview

The four corpora were compiled using Sketch Engine. Table 2 summarizes the basic statistics for the four corpora, including token counts, unique words, and type-token ratio (TTR).

Table 2. Overview of the Corpora

Corpus	Tokens	Unique Words	TTR
Human	25,653	2,741	0.107
ChatGPT-5	21,931	2,855	0.130
Gemini 2.5	26,494	3,385	0.128
Claude 4.5	27,375	3,512	0.128

Claude and Gemini produced the largest corpora and exhibited the highest lexical variety, while ChatGPT generated fewer tokens but still maintained substantial diversity. AI-generated texts show higher lexical diversity than human-authored texts, likely due to reduced repetition of pragmatic markers typical in spoken discourse. TTR is calculated by dividing the number of unique words (types) by the total number of words (tokens), and a higher TTR indicates greater lexical variety (McEnery & Hardie, 2012). Correlation analysis examined whether corpus size influenced lexical diversity. Pearson correlation revealed no significant relationship between token count and TTR ($r = -0.337$, $p = 0.663$), indicating that the higher TTR in AI-generated texts was not an artifact of corpus size differences. As expected, corpus size showed a strong positive correlation with absolute number of types ($r = 0.987$, $p = 0.013$).

4.2 Keyword analysis

Sketch Engine produced word lists for each of the four corpora. Keyword analysis revealed distinct lexical profiles across the corpora, with proper nouns (e.g., names of people, places, festivals, and language families) excluded to emphasize general patterns. A keyness score was calculated using Sketch Engine's formula:

$$\text{Score} = \frac{\text{Relative frequency (focus)} + N}{\text{Relative frequency (reference)} + N}$$

In this formula, relative frequencies are occurrences per million words, and N is a smoothing parameter (default = 1) to avoid division by zero. Higher scores indicate words more characteristic of the focus corpus relative to a general reference corpus, EnTenTen2021 (2021). Appendix A lists the top 20 keywords by keyness score for each corpus. Analysis of keyword types revealed systematic differences in discourse features. Of the human corpus's top 20 keywords, 10 (50%) were discourse markers (*hmhm, uh, um, uh-uh, uh-huh, hm-hm, mmhm, ah, yeah, mmm*), compared to two (10%) in

ChatGPT (*um, uh*), five (25%) in Gemini (*um, mmm, gosh, uh, hmm*), and four (20%) in Claude (*um, mm-hmm, uh, er*). This difference was statistically significant ($\chi^2 = 8.97$, df = 3, p = 0.030), indicating that human-authored texts contained proportionally more conversational discourse features among their most distinctive lexical items. AI corpora's top keywords included more content-focused vocabulary (*teacher-centred, pessimism, optimism, student-centred*), suggesting a shift toward conceptual rather than interactional language use. Notably, all four corpora shared certain content keywords (*phobia, superstition, fad*) due to the common topics derived from the original textbook materials (Section 3.1), indicating successful topic replication across AI-generated and human texts.

4.3 N-grams analysis

An examination of the n-grams identified frequent multi-word sequences (3-grams and 4-grams only) in each corpus. Appendix B presents the top 20 n-grams by frequency for each corpus. The human corpus displayed diverse, conversational 3- and 4-grams, including questions (*what do you, what kind of*), hedges (*a little bit*), and narrative phrases (*when I was, I thought it was*), reflecting natural spoken variability. In contrast, AI corpora—particularly ChatGPT and Claude—featured highly repetitive, formulaic sequences such as *want to talk about*, *thank you for listening*, and *let me give you*, suggesting scripted, monologue-style output typical of prompted AI responses rather than authentic dialogue. Chi-square tests revealed highly significant differences in formulaic language distribution. ChatGPT exhibited 61.3% formulaic n-grams (274 of 447 total n-gram tokens), and Claude showed 50.3% (177 of 352 tokens), compared to 0% in human texts (0 of 432 tokens) and 2.5% in Gemini (8 of 315 tokens) ($\chi^2 = 389.52$, df = 3, $p < 0.001$). Analysis of conversational versus non-conversational n-grams revealed that the human corpus contained 38.6% conversational/interactive patterns (22 of 57 unique n-grams), including question forms (what do you, do you think, what kind of), personal narratives (when I was, I thought it was), and interactive phrases (you want to, how often do you). AI corpora showed markedly reduced conversational patterns: ChatGPT (9.1%, 3 of 33), Gemini (17.5%, 7 of 40), and Claude (11.4%, 5 of 44) ($\chi^2 = 18.64$, df = 3, $p < 0.001$).

Concentration analysis examined how evenly n-grams were distributed by calculating the percentage of total n-gram tokens accounted for by the top 10 most frequent n-grams. ChatGPT's top 10 n-grams accounted for 58.0% of total usage, significantly higher than human (28.5%), Gemini (31.1%), and Claude (34.7%) corpora ($\chi^2 = 156.33$, df = 3, $p < 0.001$). This extreme concentration indicates that ChatGPT relied heavily on a small set of repetitive patterns rather than varied multi-word sequences characteristic of spontaneous speech. These statistical analyses strongly indicate that ChatGPT and Claude generate discourse characterized by scripted presentation templates rather than spontaneous conversational interaction, with Gemini performing substantially closer to human patterns in formulaic usage and n-gram diversity.

4.4 Discourse markers

A chi-square test was carried on the wordlists of all four corpora, with the human corpus containing 2.52% discourse markers compared to 1.15-1.45% in AI corpora. The analyses revealed significant differences in discourse marker usage between human and AI-generated corpora. Pairwise chi-square tests showed that the human corpus differed significantly from all three AI corpora in discourse marker frequency: Human vs. ChatGPT ($\chi^2 = 101.45$, $p < 0.001$), Human vs. Gemini ($\chi^2 = 105.82$, $p < 0.001$), and Human vs. Claude ($\chi^2 = 114.24$, $p < 0.001$). Notably, the three AI corpora did not differ

significantly from each other (all $p > 0.5$), suggesting similar limitations across models in producing discourse markers. Appendix C shows the discourse marker frequencies. These were categorized based on Carter & McCarthy (2006).

The findings suggest that current LLMs tend to generate relatively neutral, planned discourse that lacks many of the pragmatic and interactional features characteristic of spontaneous conversation. Given the importance of discourse markers, conversational n-grams, and pragmatic variability in developing EFL listening competence, these findings urge caution when using AI-generated content as primary pedagogical materials. The pedagogical implications of these differences are discussed in Section 5.

5. Discussion

The findings of this study reinforce and extend the conclusions drawn in the pilot study, which highlighted ChatGPT's limitations in replicating authentic spoken language. While the current analysis includes multiple LLMs and a larger human corpus, the core issue remains: AI-generated texts, despite their lexical richness and grammatical accuracy, lack the spontaneity, interactional nuance, and discourse-level features that characterize natural conversation. The AI-generated texts did display some formulaic language that may have benefits for students preparing to give context-governed presentations.

This absence of authenticity in the AI-generated texts has important pedagogical implications. Spoken language is inherently messy, co-constructed, and context-sensitive. Features such as hesitations, fillers, backchannels, and elliptical structures are not peripheral—they are central to how meaning is negotiated in real-time. In terms of EFL learners' ability to convey meaning when speaking in English, this may not be a significant issue. However, the lack of these features in AI-generated texts means that learners exposed primarily to AI-generated materials may miss critical opportunities to develop pragmatic competence, listening comprehension, and fluency. The tendency of LLMs to produce structured, monologic discourse with formulaic expressions strongly suggests architectural and training biases. These models are optimized for clarity, coherence, and task completion, not for replicating the unpredictability of human speech. While this makes them useful for generating clean instructional input, it limits their value for tasks that require exposure to authentic interactional patterns.

Importantly, the inclusion of Gemini 2.5 and Claude 4.5 in this study reveals substantial variation in how different LLMs approximate authentic spoken discourse. Gemini's outputs demonstrated notably greater conversational variability, with only 2.5% formulaic n-grams compared to ChatGPT's 61.3% and Claude's 50.3%, suggesting that model architecture significantly influences linguistic authenticity. This superior performance may stem from Gemini's multimodal training and 1-million-token context window, enabling better maintenance of conversational coherence. In contrast, Claude's safety-optimized architecture appears to constrain naturalistic variation in favor of clarity—a trade-off that may benefit beginner-level instruction or formal presentation training. ChatGPT's extreme reliance on formulaic sequences (58% of n-gram tokens concentrated in the top 10 patterns) suggests optimization for consistency over authenticity. These architectural biases have distinct pedagogical implications: Gemini may be better suited for intermediate-to-advanced listening practice, Claude for controlled practice emphasizing accuracy, and ChatGPT for scaffolded exercises requiring predictable patterns. However, even Gemini achieved only 17.5% conversational n-grams compared to 38.6% in human

texts, suggesting that current LLMs, regardless of their specific optimizations, face fundamental limitations in modeling spontaneous interaction. Materials developers should therefore select models strategically based on pedagogical objectives while recognizing that no current LLM can fully substitute for authentic human discourse.

For educators and materials developers, these findings underscore the need for critical engagement with AI tools. Rather than viewing LLMs as replacements for human-authored texts, they should be seen as supplementary resources. Their outputs can serve as scaffolds, templates, or controlled input, but they require adaptation, human oversight, and contextualization to meet the communicative needs of learners. Specifically, teachers might systematically inject discourse markers, hesitations, and repair sequences into AI-generated dialogues to create hybrid materials that combine AI efficiency with authentic conversational features. There may also be a case for using LLMs to create texts and manually adding features that the teacher/materials writer deems necessary. Guidelines and rubrics for assessing pragmatic features in LLM-generated materials may be helpful.

Finally, the study underscores the value of corpus-based evaluation for assessing AI-generated materials in EFL contexts. Quantitative metrics—such as lexical variety (type-token ratios), n-gram frequency, and discourse markers—offer a transparent, replicable framework for pinpointing linguistic gaps, like the scarcity of informal fillers or spontaneous chunks in ChatGPT outputs. As LLMs evolve rapidly, sustained empirical research remains crucial to promote their responsible, effective integration into language education without compromising exposure to authentic communicative features.

6. Conclusion

This study examined the linguistic characteristics of EFL listening materials generated by three leading LLMs—ChatGPT-5, Gemini 2.5, and Claude 4.5—and compared them to a human-authored corpus. Building on the pilot study, which focused solely on ChatGPT-4, this expanded analysis provides a more comprehensive view of how current AI models perform in replicating authentic spoken language. The results reveal a consistent pattern: while AI-generated texts exhibit high lexical diversity and grammatical accuracy, they lack the spontaneous, interactional features that define natural conversation. Discourse markers, conversational n-grams, and pragmatic variability were significantly underrepresented in all three AI corpora. These findings suggest that LLMs, even in their latest iterations, continue to produce language that resembles scripted monologue rather than authentic dialogue.

In response to the research questions:

> RQ1 What linguistic differences in lexical variety, n-grams, and discourse markers exist between texts generated by ChatGPT-5, Gemini 2.5, Claude 4.5, and human-authored EFL listening materials?

The linguistic differences include greater lexical variety in AI outputs but markedly reduced use of discourse markers (e.g., *well, you know, like*) and conversational n-grams (e.g., *a little bit, on the other hand*). These patterns reflect a shift toward planned, presentation-style language in AI-generated materials, which contrasts sharply with the messy, incremental, and interactive nature of human speech. Gemini showed relatively more conversational variability, but still fell short of human-authored norms.

> RQ2 What are the implications of these differences for EFL materials development?

The pedagogical implications are clear. While AI-generated texts may be suitable for structured tasks, controlled practice, or reading-based input, they are less effective for developing listening skills and pragmatic competence. Educators should use these tools judiciously, supplementing them with authentic materials and adapting AI outputs—through editing or targeted prompting—to include more naturalistic features. Hybrid approaches may offer a practical solution, combining AI efficiency with human oversight to enhance authenticity.

Prompt design plays a key role in this process. Although this study employed structured prompts to elicit conversational elements, the results indicate that even carefully crafted input cannot fully overcome the models' inherent biases toward formal, monologic language. Future work should explore advanced prompting strategies, including iterative refinement, role-playing constraints, and multimodal conditioning (e.g., incorporating prosodic cues or interactional context), to better approximate authentic speech.

At the same time, this study is limited by its reliance on text-based output rather than synthesized speech. Follow-up studies could evaluate how learners process prosodically enriched AI-generated audio, or how exposure to hybridized human-AI listening input affects comprehension and pragmatic awareness in classroom settings.

As the boundaries between human and machine-generated content continue to blur, educators and researchers must critically examine not just what AI can do, but how and why its linguistic limitations matter in language learning contexts. Addressing these gaps will likely require collaboration between applied linguists, educators, and AI developers to co-design more dialogically sensitive prompting frameworks and evaluation tools. Only then can we begin to harness the full pedagogical potential of LLMs without sacrificing the subtle, emergent features that make human communication meaningful.

LLMs offer valuable support for language education, but their current limitations necessitate a balanced, evidence-informed approach. By combining AI-generated content with human-authored materials and applying corpus-based evaluation methods, educators can ensure that learners receive linguistically rich, pedagogically sound input that supports real-world communicative competence. A thoughtful integration of both AI and human resources can help bridge the authenticity gap while retaining the efficiency and adaptability that make LLMs attractive tools for language educators.

Declarations and Acknowledgment:

The author declares no conflicts of interest. This pilot study that precedes this paper was first presented at the *International Conference on Globalisation/Deglobalisation in Languages, Education, Culture and Communication (GLECC2025).*

References

Adolphs, S., & Knight, D. (2010). Building a spoken corpus: What are the basics? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 38–52). Routledge.

Anthropic. (2025, September 29). *Introducing Claude Sonnet 4.5*. Anthropic. https://www.anthropic.com/news/claude-sonnet-4-5

Bewersdorff, A., Hartmann, C., Hornberger, M., Seßler, K., Bannert, M., Kasneci, E., Kasneci, G., Zhai, X., & Nerdel, C. (2025). Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education. *Learning and Individual Differences, 118*, 102601. https://doi.org/10.1016/j.lindif.2024.102601

Biber, D., & Conrad, S. (2019). *Register, genre, and style* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108686136

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* Longman.

Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English*. Cambridge University Press.

Carter, R., & McCarthy, M. (2017). Spoken grammar: Where are we and where are we going? *Applied Linguistics, 38*(1), 1–20. https://doi.org/10.1093/applin/amu080

Chen, X., Zou, D., Xie, H., Cheng, G., & Liu, C. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments, 11*(1), Article 26. https://doi.org/10.1186/s40561-024-00316-7

Crosthwaite, P., & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics, 3*(3), 100066. https://doi.org/10.1016/j.acorp.2023.100066

Donnellan, M. (2025). Evaluating the linguistic range of ChatGPT: A Corpus Analysis. *Proceedings of the International Conference on Globalisation in Languages, Education, Culture and Communication (GLECC 2025)* (pp. 80–86). Manchester, UK: AT Publishing. https://glecc.org/2025/wp-content/uploads/2025/09/GLECC2025-conference-proceedings-final.pdf

Fleisig, E., Smith, G., Bossi, M., Rustagi, I., Yin, X., & Klein, D. (2024). *Linguistic bias in ChatGPT: Language models reinforce dialect discrimination* (arXiv preprint arXiv:2406.08818). https://doi.org/10.48550/arXiv.2406.08818

Flowerdew, L. (2004). The argument for using English specialized corpora to understand academic and professional language. In U. Connor & T. A. Upton (Eds.), *Discourse in the professions: Perspectives from corpus linguistics* (pp. 11–33). John Benjamins. https://doi.org/10.1075/scl.16.02flo

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning, 67*(S1), 155–179. https://doi.org/10.1111/lang.12225

Goh, C. C. M., & Aryadoust, V. (2025). Developing and assessing second language listening and speaking: Does AI make it better? *Annual Review of Applied Linguistics, 45*, 179–199. https://doi.org/10.1017/S0267190525100111

Google DeepMind. (2025). *Introducing Gemini 2.5*. Google. https://deepmind.google/technologies/gemini/

Google Workspace. (2025, June). Gemini in Google Classroom is now available to all Google Workspace for Education editions, with added features. *Google Workspace Updates*. https://workspaceupdates.googleblog.com/2025/06/gemini-google-classroom-all-edu-editions.html

Gries, S. T. (2021). *Statistics for linguistics with R: A practical introduction* (3rd ed.). De Gruyter Mouton. https://doi.org/10.1515/9783110718256

Hajikhani, A. (2024). A critical review of large language models: Sensitivity, bias, and the path toward specialized AI. *Quantitative Science and Technology Studies, 5*(3), 736–756. https://doi.org/10.1162/qss_a_00310

Hochmair, H. H., Juhász, L., & Kemp, T. (2024). Correctness comparison of ChatGPT-4, Gemini, Claude-3, and copilot for spatial tasks. *Transactions in GIS, 28*(7), 2219–2231. https://doi.org/10.1111/tgis.13233

Huang, H., Zheng, O., Wang, D., Yin, J., Wang, Z., Ding, S., Yin, H., Xu, C., Yang, R., Zheng, Q., Pu, B., Tseng, H.-C., Bu, S. J., Yu, S., Zhu, W., Cai, W., Wei, X., Gao, Y., Jin, B., ... Xu, C. (2023). ChatGPT for shaping the future of dentistry: The potential of multi-modal large language model. *International Journal of Oral Science, 15*(1), Article 29. https://doi.org/10.1038/s41368-023-00239-y

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G. (2023). *ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography, 1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Koester, A. (2010a). Building small specialised corpora. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 66–79). Routledge.

Koester, A. (2010b). *Workplace discourse*. Continuum.

Leech, G. (2000). Grammars of spoken English: New outcomes of corpus-oriented research. *Language Learning, 50*(4), 675–724. https://doi.org/10.1111/0023-8333.00143

Lexical Computing CZ, s.r.o. (2021). *enTenTen21: English Web Corpus 2021*. Sketch Engine. https://www.sketchengine.eu/ententen-english-corpus/

Liu, M., Okuhara, T., Dai, Z., Huang, W., Gu, L., Okada, H., Furukawa, E., & Kiuchi, T. (2025). Evaluating the effectiveness of advanced large language models in medical knowledge: A comparative study using Japanese national medical examination. *International Journal of Medical Informatics, 193*, 105673. https://doi.org/10.1016/j.ijmedinf.2024.105673

Liu, Z.-M., Hwang, G.-J., Chen, C.-Q., Chen, X.-D., & Ye, X.-D. (2024). Integrating large language models into EFL writing instruction: Effects on performance, self-regulated learning strategies, and motivation. *Computer Assisted Language Learning.* Advance online publication. https://doi.org/10.1080/09588221.2024.2389923

McCarthy, M., & Carter, R. (2014). *Spoken language and applied linguistics*. Cambridge University Press.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice.* Cambridge University Press.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics* (2nd ed.). Edinburgh University Press.

Meta AI. (2025, April 5). *The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation.* https://ai.meta.com/blog/llama-4-multimodal-intelligence/

Microsoft. (2025). *Microsoft Copilot* [Large language model]. Microsoft. https://copilot.microsoft.com/

Nelson, M. (2010). Building a written corpus: What are the basics? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 53–65). Routledge.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom.* Cambridge University Press.

OpenAI. (2025, August 7). *Introducing GPT-5*. OpenAI. https://openai.com/gpt-5

Römer, U. (2020). Corpus linguistics and language pedagogy: The state of the art and beyond. In V. X. Wang (Ed.), *Handbook of research on teaching and learning in K-12 education* (pp. 89–106). IGI Global. https://doi.org/10.4018/978-1-7998-0951-7.ch006

Sandler, M., Choung, H., Ross, A., & David, P. (2024). *A linguistic comparison between human and ChatGPT-generated conversations* (arXiv preprint arXiv:2401.16587). https://doi.org/10.48550/arXiv.2401.16587

Schillinger, P., Vogel, J., Vlasov, M., Ewerton, M., Gams, A., Hitzler, R., Mattamala, M., Gutierrez, C., & Falco, F. (2024). Capabilities of large language models in control engineering: A benchmark study on GPT-4, Claude 3 Opus, and Gemini 1.0 *Ultra. arXiv*. https://doi.org/10.48550/arXiv.2404.03647

Seo, H., Hwang, T., Jung, J., Kang, H., Namgoong, H., Lee, Y., & Jung, S. (2025). Large Language Models as Evaluators in Education: Verification of Feedback Consistency and Accuracy. *Applied Sciences, 15*(2), 671. https://doi.org/10.3390/app15020671

Tao, H. (2003). Turn initiators in spoken English. *Journal of Pragmatics, 35*(1), 25–42.

Wei, L. (2023). Artificial intelligence in language instruction: Impact on English learning achievement, L2 motivation, and self-regulated learning. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1261955

Willis, D., & Willis, J. (2007). *Doing task-based teaching*. Oxford University Press.

xAI. (2025, February 19). *Grok 3 Beta — The age of reasoning agents*. https://x.ai/news/grok-3

Yilmaz, R., Yilmaz, F. G. K., & Keser, H. (2024). Have we reached AGI? Comparing ChatGPT, Claude, and Gemini to human literacy and education benchmarks. *arXiv.* https://doi.org/10.48550/arXiv.2407.09573

Zheng, Y., Koh, H. Y., Ju, J., Nguyen, A. T. N., May, L. T., Webb, G. I., & Pan, S. (2025). Large language models for scientific discovery in molecular property prediction. *Nature Machine Intelligence, 7*(3), 437–447. https://doi.org/10.1038/s42256-025-00994-z

About the Author:

Mark Donnellan is a lecturer in the Faculty of Informatics at Kindai University in Osaka, Japan. His research interests include task-based language teaching, corpus linguistics, language produced by generative AI, and integrating technology in the EFL classroom. He also has extensive experience organizing virtual exchanges between Japanese university students and students at various universities in Europe.

Appendix A

Top 20 Keywords from each Corpus

Rank	Human Corpus Keyword (Score)	ChatGPT Corpus Keyword (Score)	Gemini Corpus Keyword (Score)	Claude Corpus Keyword (Score)
1	hmhm (350.8)	um (422.2)	um (458.9)	phobia (363.8)
2	phobia (291.3)	phobia (227.2)	mmm (416.4)	um (309.7)
3	uh (255.8)	teacher-centred (181.4)	gosh (213.4)	fad (187.6)
4	um (247.9)	fad (175.6)	phobia (156.9)	mm-hmm (170.1)
5	uh-uh (221.3)	student-centred (169.4)	fremdschämen (114.1)	student-centered (159.5)
6	uh-huh (206.3)	superstition (164.0)	teacher-centered (110.3)	suggestopedia (146.4)
7	hm-hm (195.7)	pessimism (94.9)	fad (109.2)	superstition (142.4)
8	mmhm (155.5)	fremdschämen (92.1)	student-centered (82.8)	teacher-centered (142.0)
9	ah (151.8)	suggestopedia (91.8)	superstition (79.4)	uh (117.9)
10	yeah (136.2)	solar-powered (58.8)	uh (69.1)	russian-language (66.4)
11	superstitious (118.4)	optimism (57.7)	counter-intuitive (52.1)	worrier (65.2)
12	prepone (117.3)	uh (54.0)	fluency (43.3)	divisor (62.9)
13	superstition (117.0)	turmeric (52.8)	pessimism (39.6)	optimistic (62.9)
14	fad (100.2)	hello (48.6)	re-trains (38.7)	egghead (62.7)
15	okay (88.2)	laughing (47.1)	clothing-related (38.6)	unlucky (58.6)
16	fremdschämen (78.9)	lower-gdp (46.6)	hmm (38.6)	anti-theft (57.5)
17	demotivating (74.9)	schämen (46.6)	trypanophobia (38.6)	pessimism (57.3)
18	curcumin (72.5)	half-looking (46.6)	half-listening (38.1)	businesspeople (52.9)
19	huh (66.7)	pink-themed (46.4)	crisis-affected (37.8)	er (51.6)
20	mmm (61.8)	prepone (46.3)	often-quoted (37.8)	motivator (51.3)

Appendix B

Top 20 N-grams in each Corpus

Rank	Human Corpus N-gram (Freq.)	ChatGPT Corpus N-gram (Freq.)	Gemini Corpus N-gram (Freq.)	Claude Corpus N-gram (Freq.)
1	around the world (16)	to talk about (29)	in the world (11)	Thank you for (14)
2	in the world (15)	want to talk (27)	you have a (11)	Thank you for listening (13)
3	a little bit (10)	want to talk about (27)	do you think (9)	you for listening (13)
4	the number of (8)	I want to talk (27)	I think it (8)	Let me give you (10)
5	in New Zealand (8)	Thank you for (24)	all over the (8)	around the world (10)
6	what do you (8)	Thank you for listening (24)	all the time (8)	Let me give (10)
7	what kind of (8)	you for listening (24)	all over the world (8)	me give you (10)
8	when I was (8)	today I want (10)	over the world (8)	to talk about (9)
9	a long time (8)	today I want to (10)	Thank you for (8)	tell you about (9)
10	I don't know (7)	I liked how (9)	I think I (8)	the United States (8)
11	part of the (7)	around the world (9)	it is a (8)	What kind of (7)
12	the United States (7)	a lot of (8)	to be honest (7)	What about you (7)
13	the other hand (7)	a long time (7)	is a very (7)	How did you (7)
14	On the other (6)	It made me (6)	around the world (7)	one of the (7)
15	to go to (6)	helps us understand (6)	part of the (7)	can be very (6)
16	go to the (6)	in the world (6)	when I was (7)	did you do (6)
17	On the other hand (6)	it is not (6)	Do you have (7)	Today I want (6)
18	I thought it (6)	I did not (6)	number of countries (5)	Today's topic is (6)
19	thought it was (6)	to see the (6)	have a lot (5)	Today I want to (6)
20	I thought it was (6)	did you finish (6)	have a lot of (5)	you want to (6)

Appendix C

Discourse Marker Frequencies

Marker	Type	Human	ChatGPT	Gemini	Claude
um	Hesitation	57	83	109	76
uh	Hesitation	61	11	17	30
er	Hesitation	0	0	0	17
yeah	Response	173	109	18	47
oh	Response	74	63	40	31
ah	Response	63	0	10	0
well	Response	80	31	78	100
okay	Response	84	8	0	0
uh-huh	Backchannel	7	0	0	0
hmhm	Backchannel	9	0	0	0
mm-hmm	Backchannel	0	0	0	5
hm-hm	Backchannel	5	0	0	0
mmhm	Backchannel	4	0	0	0
hmm	Acknowledgment	7	10	7	8
mmm	Acknowledgment	4	4	28	2
huh	Acknowledgment	12	0	0	0
gosh	Acknowledgment	0	0	18	0
TOTAL		**646**	**319**	**325**	**316**
% of tokens		**2.52%**	**1.45%**	**1.23%**	**1.15%**
Unique types		**15**	**6**	**7**	**7**

***Markers categorized following Carter & McCarthy (2006)