Scientific Research Publishing

Open Journal of Modern Linguistics, 2025, 15(5), 754-772
https://www.scirp.org/journal/ojml
ISSN Online: 2164-2834
ISSN Print: 2164-2818

The Use of Hedging Devices and Engagement Markers in AI-Generated and Human-Written Essays: A Corpus-Based Comparison

Naif Almulla

Department of English Language, Majmaah University, Majmaah, Saudi Arabia
Email: n.almulla@mu.edu.sa

How to cite this paper: Almulla, N. (2025). The Use of Hedging Devices and Engagement Markers in AI-Generated and Human-Written Essays: A Corpus-Based Comparison. *Open Journal of Modern Linguistics, 15,* 754-772.
https://doi.org/10.4236/ojml.2025.155044

Received: August 3, 2025
Accepted: September 15, 2025
Published: September 18, 2025

Copyright © 2025 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

AI tools, such as large language model (LLM) chatbots, have made attaining many learning and educational tasks much easier, and probably even more efficient. These easily accessible and useful resources have made it easy for many students to rely on these tools. This reality has raised some concerns among educators regarding possible negative effects on cognitive abilities and ethical considerations that might result from extensive use of such tools. In order to address this issue, some researchers have attempted to see if LLM chatbots have a unique style that educators can identify and distinguish from the human writing style. Contributing to this effort, this corpus-based study aims to examine two characteristics found in academic writing (i.e., "hedging" and "engagement markers") to see whether there are differences between AI-generated text and human-written text in the use of these two. In order to analyze and compare the two types of text for hedging and engagement markers, two main text collections were compiled, each consisting of about 20,000 words. Taking a mainly descriptive statistics approach to analyze the data, the results showed that AI-generated text used noticeably more hedging words, whereas human-written text used noticeably more engagement markers. These findings corroborate previous research, provide a better conceptualization of the issue, and emphasize the need for further research on the topic.

Keywords

Artificial Intelligence (AI), LLMs, Academic Writing, Essay Writing, Hedging, Engagement, Corpus-Based

1. Introduction

During the past decade, artificial intelligence (AI) tools have offered a wide range

of functions as learning aids that students find valuable. One notable advantage of these tools is their ability to analyze and generate large amounts of text. As more students rely on AI-generated content, an important question emerges: what kind of language does AI produce, and how does it differ from human language?

To the best of the writer's knowledge, and despite the significance of this issue, research studies that look specifically into distinguishing features to differentiate between AI-generated and human-written text styles are very few. The focus of research on this topic centralizes more dissimilarities between the two types of texts because similarities are already assumed to exist, since AI chatbots are trained on natural human language data and are very accurate in mirroring human language styles. A large number of studies have explored the role of AI in higher education (Zawacki-Richter, Marín, Bond, & Gouverneur, 2019; Zhai et al., 2021), but few have examined the linguistic characteristics of AI-generated language versus natural human language (e.g., AlAfnan & MohdZuki, 2023; Jiang & Hyland, 2024a, 2024b; Zhang & Crosthwaite, 2025).

Large language models (LLMs), like ChatGPT, are based on algorithms that are learned from extensive datasets composed of internet texts (Jiang & Hyland, 2024b). They are designed to detect and learn from statistical patterns and relationships between words and phrases found in large bodies of texts (Jiang & Hyland, 2024b). This advanced statistical learning from language data marks a significant advancement in natural language processing (NLP) (Kasneci et al., 2023; Jiang & Hyland, 2024b). The recent revolution in AI with regard to text processing has proven beneficial for many tasks. However, due to the inherent limitations of AI design, AI bots may exhibit shortcomings such as bias and hallucinations (Kasneci et al., 2023). Bias can occur when models are trained on datasets that contain underlying biases (Kasneci et al., 2023). According to Kasneci et al. (2023), a pressing concern in educational and academic contexts is the "difficulty to distinguish model-generated from student-generated answers" (Kasneci et al., 2023: p. 8).

At the same time, since AI assistant bots have only recently become publicly available, they have already facilitated the achievement of a variety of tasks. In higher education, they have supported instructors by assisting with grading large volumes of essay assignments and offering necessary scaffolding and student-specific feedback (Zhang & Crosthwaite, 2025). Another major concern involves plagiarism and copyright issues (Kasneci et al., 2023). Because these models are trained on large datasets, they may occasionally generate similar or even identical content.

2. Literature Review

2.1. Large Language Models and Writing

Most research on AI large language models (LLMs) in educational contexts has focused on their capabilities and applications in teaching (Casal & Kessler, 2023). However, over the past two years, researchers have begun to examine the potential negative effects these tools may have on students' cognitive abilities, such as prob-

lem-solving and critical thinking (Casal & Kessler, 2023; Kasneci et al., 2023). Beyond the concern that AI might impair such cognitive skills, an emerging issue involves academic integrity and plagiarism. This has led to growing interest in how AI-generated text can be distinguished from human-written text (AlAfnan & MohdZuki, 2023; Casal & Kessler, 2023; Jiang & Hyland, 2024b; Kasneci et al., 2023; Zhang & Crosthwaite, 2025). This becomes especially important because online originality detection tools overall fail to render accurate detection rates (AlAfnan & MohdZuki, 2023), which, if relied on by instructors, may lead to unfair penalties on students. Based on several observed differences between the two, Zhang and Crosthwaite (2025) suggest that educators should recognize their students' individual experiences and social backgrounds in order to better detect distinctions between AI-generated and human-written essays.

In a study exploring whether applied linguistics journal reviewers could differentiate between AI-generated and human-written abstracts, Casal and Kessler (2023) collected data through a survey of 72 reviewers and semi-structured interviews with 28 journal reviewers. The participants included 12 assistant professors, 17 associate professors, 6 full professors, and 22 whose academic ranks were not specified. They represented diverse research areas within applied linguistics, including second language acquisition, corpus linguistics, psycholinguistics, L2 vocabulary, and L2 writing.

The results revealed that reviewers were able to correctly identify whether an abstract was AI-generated or human-written only 39% of the time. They were generally more successful at identifying human-written abstracts than AI-produced ones. Additionally, "specificity", "vagueness", and "lack of details" were frequently reported by reviewers as markers of AI-generated texts. However, despite identifying these traits, reviewers were able to apply this criterion only 29% of the time.

AlAfnan and MohdZuki (2023) conducted a stylistic study aimed at identifying features of texts produced by ChatGPT. One of the primary concerns motivating the study was the inability of many AI plagiarism checkers to provide accurate reports. The study examined several features (e.g., sentence length, paragraph structure, word choice, pronouns, lexical density, etc.) in an attempt to see if AI-produced text has a writing style. The findings showed that "Academic Writing" responses were composed of paragraphs consisting of 2 - 4 sentences, and sentences were composed of words ranging from 16 to 19 words. The responses were mainly generated in the declarative mood, and the lexical density in "academic writing" responses was "high".

2.2. Previous Studies

One recent study that compared AI-generated essays to human-written essays in academic discourse is Jiang and Hyland (2024b). In this study, the researchers analyzed 145 argumentative essays generated by ChatGPT 4.0 (over 72,000 words) on various topics and compared them to 145 essays written by British university stu-

dents (over 78,000 words) to examine differences in "reader engagement" markers, based on Hyland's (2005) Model of Interaction and Engagement. According to this model, academic writers actively engage readers (using five key techniques: reader mentions, questions, appeals to shared knowledge, directives, and personal asides).

Before analyzing the full dataset, the two researchers independently coded a random sample and resolved disagreements in an attempt to ensure a high level of interrater reliability. When comparing the two corpora (i.e., AI-generated essays and human-written essays) for engagement markers, they found a significant difference: AI-generated essays used far fewer engagement markers (393) compared to student-written essays (1326).

This finding aligns with previous work by the same authors examining lexical bundle usage in AI-generated versus human-written text (Jiang & Hyland, 2024a). That earlier study found that ChatGPT-generated essays exhibited "rigid and formulaic patterns", suggesting that AI tools may have limited creativity in language production. Together, these findings point to distinctive stylistic features of AI-generated texts.

In a fine-grained study comparing academic essays produced by ChatGPT and human English L2 writers, Zhang and Crosthwaite (2025) sought to examine differences in the use of "lexical items" and "collocations" between the two types of texts. The 30 argumentative essays written by EFL learners were final drafts submitted for an academic writing course on various controversial topics. Students had received feedback on earlier drafts to help improve their critical thinking and proper use of evidence and citations. The learners came from diverse L1 backgrounds, with proficiency levels ranging from upper-intermediate to advanced, and were enrolled at an Australian university.

Another 30 argumentative essays were generated by ChatGPT using a prompt that included instructions specifying certain requirements such as a word limit of 1,000 words, the use of APA 7th edition citations, and no subheadings (commonly found in ChatGPT-produced text). The data were cleaned to ensure suitability for the purposes of this comparative study.

With regard to lexical items, the findings revealed that ChatGPT produced significantly more "diverse" and "formal" vocabulary. In contrast, the EFL learners tended to use "simple language", despite their relatively high proficiency levels. When comparing collocations, ChatGPT's essays were found to focus mostly on "contemporary" and "forward-looking issues", which reflected a sense of "sophistication". On the other hand, the human L2 writers focused more on topics such as "leadership", "work-related issues", and "environmental concerns" (themes that are more directly connected to lived experiences and societal contexts).

2.3. Hedging in Academic Writing

Like any other register of language with its own identity and features, academic writing can be distinguished from other forms of writing by its own characteris-

tics. It is primarily represented by research articles, in which scholars "*report*", "*argue*", or "*describe*" their findings and claims (Hyland, 2005). According to Hyland (1996), when making claims about their research, scientists need to balance between *asserting their findings with confidence*—based on which they gain credibility—and *ensuring their claims are not perceived as overstatements*. As Hyland (1996) notes, "only claims that seem to be objectively legitimate and sensitive to audience expectations are likely to be ratified" (p. 255).

Although researchers' arguments and claims are typically propositional and impersonal, they still need to make linguistic choices the audience finds persuasive in order to get their readers to agree with what they are proposing (Hyland, 2005). While the use of hedges (e.g., *may*, *could*, *perhaps*, *probably*, etc.) was once discouraged and viewed negatively in academic writing due to its association with "subjectivity", this perception has shifted over the past few decades, as academic writers have increasingly sought to engage with their readers (Hyland, 2005). Since writers need to meet content-based functions (i.e., *adequacy of claims*) and reader-based functions (i.e., *acceptability of claims*), hedging devices are now recognized as important because of the simultaneous "polypragmatic" communicative functions they convey (Hyland, 1996).

The absence of hedging strategies and devices can lead to overly definitive statements, an issue that is particularly important for novice writers, such as English language learners (Hyland, 1996). This is "critical", according to Hyland (1996), because if learners fail to hedge statements appropriately, it may prevent them from effectively participating in "a research world" that is largely dominated by English. This concern is especially relevant in the context of English for Academic Purposes (EAP), a branch of the broader field of English for Specific Purposes (ESP). In many universities around the world, students from non-English-speaking backgrounds are required to learn English in order to pursue higher education degrees. As such, familiarity with the conventions of academic writing, especially the accurate use of hedging, is vital in academic writing (Hyland, 1996).

2.4. The Present Study

Many studies have investigated how AI bots can be used as a learning aid in higher education (e.g., Kim & Adlof, 2024; Farrokhnia et al., 2024); some have focused on their capabilities in teaching and learning writing (e.g., Lu et al., 2024), some on their role in writing (e.g., Alshalan & Alyousef, 2024), and some others have focused on their ability to assess AI-generated essays (e.g., Mizumoto & Eguchi, 2023). However, there is still a very small number of studies that have examined how comparable AI-generated texts are to human-written texts. In addition, as Jiang and Hyland (2024b) highlight, with the growing influence of AI bots on academic writing and the many studies that have been conducted on AI in relation to language, studies looking specifically at how similar AI-generated texts are to human writing and whether they have the same degree of reader engagement—which is an important aspect of academic writing—are still lacking.

For these reasons, the present study aims to investigate the topic and examine what differences, if any, occur between ChatGPT-generated language and human academic writing in using hedging words/phrases. This study opts to work with ChatGPT-4o (a more advanced version of ChatGPT) since it is purported by OpenAI (2023) to produce "more coherent and contextually appropriate responses" as well as "more human-like conversations". Focusing on "hedging" devices as markers of engagement between writer and reader, this study aims to answer the following questions:

1) Are there differences between AI-generated argumentative essays and human-written argumentative essays with respect to the use of "*hedges*"?

2) Are there differences between AI-generated argumentative essays and human-written argumentative essays with respect to the use of "*engagement markers*"?

3) If there are differences between the AI-generated argumentative essays and human-written argumentative essays in the use of "*hedges*" or "*engagement markers*", are these differences statistically significant?

3. Method

By focusing on "hedging devices" and writer-reader "engagement markers", this study aims to find out: 1) whether there are differences between AI-generated (ChatGPT-4o) argumentative essays and human-written (university students) argumentative essays in the use of hedging devices, 2) which of the two types of text shows more writer-reader engagement (as indicated by increased usage of engagement markers), and 3) whether any differences between the two, if any, are statistically significant.

3.1. Data

Four argumentative essay corpora are compiled in order to investigate the objectives of the present study, which aims to compare AI and human writers' use of *hedges* and *engagement markers* in academic writing. Brief descriptions of the four text collections are provided in Table 1 for each corpus.

Table 1. Description of the four corpora used here.

	Type of Corpus	Description	No. of Words
1	AI-Generated (Default)	ChatGPT-4o argumentative essays using default prompts	10,821
2	AI-Generated (Humanized)	ChatGPT-4o argumentative essays with "humanizing" prompts	10,311
3	Human-Hand-Written	US university student essays from the LOCNESS corpus	10,214
4	Human-Typed	University student essays/personal statements from the Essay Forum website	11,100

These four corpora (each consisting of approximately 10,000 words) form two main corpora (each consisting of about 20,000 words). The two main corpora represent AI-generated essays (AIGEs) and human-written essays (HWEs). Below is an overall layout of the steps followed in composing these four corpora.

1. The AI-Generated (Default) Corpus:

For the first corpus, ChatGPT-4o was asked to generate argumentative essays based on prompts on controversial topics from the New York Times prompt list[1]. All of the chosen prompts shared the characteristic of thematic generality (i.e., they were general topics that are not specific to the US context). The prompts were input into ChatGPT-4o in the following format:

"*With no less than* 500 *words, respond to the following prompt*: '*… ESSAY PROMPT HERE…*'".

The "with no less than 500 words" directive phrase was included following the conventionally used minimum number of words in argumentative essay writing. The target word count was 10,000 words, and was achieved by the 18th essay generated by ChatGPT-4o.

2. The AI-Generated (Humanized) Corpus:

The same steps used for the AI-Generated (Default) essay corpus were followed in composing this corpus. The only difference is the addition of the "*making it as much human-like as possible*" directive phrase at the beginning of the prompt. ChatGPT-4o is considered to produce more human-like language—compared to ChatGPT-4o mini, for example. The text that was generated showed noticeable features of texts produced by human writers. The purpose of this was to see whether there were standout differences in the text produced if such prompts were used. In other words, the aim was to ensure there are no differences that can be attributed to prompt format rather than to AI style (which is the main factor under examination). The prompt used for this corpus was:

"*Making it as much human-like as possible, and with no less than* 500 *words, respond to the following prompt*: '*… ESSAY PROMPT HERE…*'".

3. The Human (Handwritten) Corpus:

The third corpus is taken from the "Louvain Corpus of Native English Essays" (LOCNESS)[2]. This is the same corpus resource used in Jiang and Hyland (2024b). However, whereas in that study, *British* university students' essays were analyzed, *American* university students' essays are used in this study. From this corpus of essays, the first 10,000 words were extracted. The purpose of adding this type of text was to make it part of the main *Human-written* corpus since it is composed of argumentative essays written by university students in a "class setting". Essays in this corpus are relatively old (written in 1995) compared to the other type of human-written essays (i.e., Typed) described below.

4. The Human (Typed) Corpus:

[1]401 Prompts for Argumentative Writing, accessed March 8th, 2025: https://static01.nyt.com/images/blogs/learning/pdf/2017/401PromptsArgumentativeWriting.pdf.
[2]https://www.learnercorpusassociation.org/, accessed March 8th, 2025.

The fourth corpus is composed of argumentative essays and personal statements found on the *Essay Forum* website[3]. Students post essay drafts on this website and receive feedback from writing experts on how to improve their essays. Although identifying the writers as native or non-native speakers may not have been attainable—since it is an online source—these essays were obtained from the *graduate* section of the website, which, to a large extent, should ensure the academic style of the text. The labels for the two types of HWEs (i.e., *Hand-written* & *Typed*) are not intended to indicate the nature of how the essays were written as much as to refer to the period during which they were written (in 1995, "*less recent*", and 2020-2025, "*more recent*").

If the claim that academic writing in the past exhibited a greater tendency towards objectivity and less personal style is true, then the Human-Hand-written essays (less recent) would show more similarities with AIGEs than the Human-Written (Typed) essays (more recent) would. However, Hyland and Jiang (2017) note that while this may have been a general idea that is usually accepted and assumed by many, it has not been examined enough. In particular, while "impersonality" may exist as a feature in academic discourse, it is not to the extent it may have been claimed to be (Hyland, 2002). In addition, while it may be used in academic writing, the use of "impersonality" certainly varies from discipline to discipline (Hyland, 2002). As shown in **Table 2**, while the use of 1st person pronouns in Human-Hand-written essays is 0.6 (closer to both AI groups of essays), it is 7.5 for Human-Written (Typed), which shows that students might be more flexible now when it comes to the use of 1st pronouns in academic writing.

Table 2. Descriptive information of the four basic corpora.

		AI-Generated (Default)	AI-Generated (Humanized)	Human-Written (Hand-Written)	Human-Written (Typed)
		No. of Words	No. of Words	No. of Words	No. of Words
Part of Speech (POS)	Nouns	3452	3098	2535	2597
	Verbs	1458	1439	1173	1602
	Adjectives	1224	1176	930	893
	Adverbs	526	502	529	558
Average Word Length		5.91	5.69	4.8	4.9
Lexical Density		0.53	0.51	0.45	0.45
Academic Words		15.92%	13.85%	9.11%	8.61%
1st Person (e.g., "I")		0.1	0.2	0.6	7.5
3rd Person (e.g., "they")		4.1	5.4	7.1	3.4

Lexically, however, based on the "lexical density" and "average word length" rates (shown in **Table 2**), Human-Hand-written and Human-written (Typed) show

[3]https://www.essayforum.com, accessed March 8th, 2025.

identical or very similar rates (0.45 and 0.45) and (4.8 and 4.9), respectively. In relation to this context, Zhou et al. (2023) investigated biology letters (which is a generally recommended field by Hyland (1996) for examining academic writing) published in the popular science journal "Nature" across a 100-year period from 1929 to 2019. Zhou et al. (2023) found that while there was a noticeable decline in "lexical density" from 2009 to 2019 (to which the period from 1939 to 1949 also showed a similar pattern), the overall trend across the 100-year period was an increase in the use of content words, which carry more semantic denotations than function words (which, in turn, results in increasing the lexical density of a text). With regard to the use of "academic words", it was found that there was a significant moderate decrease from 1929 to 2019.

Using the UAM Corpus Tool (O'Donnell, 2018), overall feature descriptions of the four corpora have been obtained and are shown in **Table 2**.

3.2. Data Analysis

In order to answer the first research question about whether there are differences between AIGEs (both *Default* & *Humanized*) and HWEs (both *Hand-written* & *Typed*) with regard to the use of hedging in academic writing, a relatively representative sample of hedging words (mostly the ones used in Adrian & Fajri, 2023) from different word classes (i.e., *nouns*, *verbs*, *adjectives*, *adverbs*, & *modals*) were searched for across the four texts. The corpus software AntConc (Anthony, 2022) was used in performing this task. The hedging words used for this search are shown in **Table 3**.

Table 3. Hedging words searched for in the four texts, adapted (with some changes) from Adrian and Fajri (2023).

Part of Speech (POS)	Targeted Words
Nouns	Assumption, implication, possibility, prediction, probability, tendency
Verbs	Appear, argue, assume, indicate, predict, propose, report, seem, suggest, tend
Adjectives	Plausible, possible, probable, rare, unlikely, likely
Adverbs	About, almost, around, generally, likely, approximately
Modals	Can, could, may, might, should, will, would

The number of instances of each of the above words was counted for each of the four corpora separately. The values for the two main corpora (AI-Generated & Human-Written) were calculated from the four sub-corpora. For example, the raw frequency of the modal "*can*" in the AI-Generated (Default) was 166, and for the AI-Generated (Humanized), it was 175. By adding these two numbers and dividing them by two, the overall AI-Generated value 170.5 is derived. In the same way, values of the overall Human-Written corpus were calculated by adding fre-

quency counts of each hedging item in the two sub-corpora (Human-Hand-written & Human-Typed) and dividing the resulting sum by two. Frequencies of the two main corpora are most essential to this study, whose results may indicate the need to further examine the four sub-corpora results.

In order to answer the second research question about which type of text (i.e., AIGEs (*Default* & *Humanized*) or HWEs (*Hand-written* & *Typed*)) shows more writer engagement with the reader through engagement markers, the "*Engagement*" component of the *Model of Interaction and Engagement* (Hyland, 2005) was applied. Using the same words as found in Jiang and Hyland (2024b), the four corpora were searched to see how engagement markers used in these texts might be classified and explained through this model. The full version of the model is represented in Figure 1.

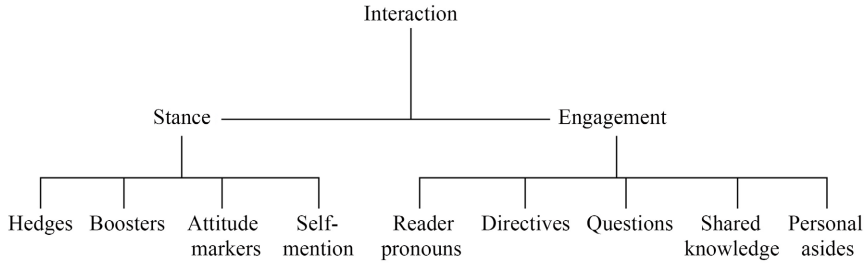

Figure 1. The model of interaction and engagement, adapted from Hyland (2005: p. 177).

The five categories of engagement are generally described in Jiang and Hyland (2024b) as:

1. Reader mentions (e.g., *we*, *let's*, *the reader*).

2. Questions, which engage the reader, such as "*Can we expect a scientist to bear this additional burden for the whole world?*" (Jiang & Hyland, 2024b).

3. Appeals to shared knowledge (e.g., *traditionally*, *apparently*).

4. Directives (e.g., *note*, *should*, *it's important to understand*).

5. Personal asides, which are brief interjections where the writer speaks directly to the reader to share a personal thought (e.g., *by the way*, *incidentally*).

The corpus software AntConc (Anthony, 2022) was used to obtain raw frequency counts of these words as they occur in the four corpora. In the case of standout trends, further qualitative analysis of these instances was conducted to see how they could be explained in terms of AI-produced language as opposed to human-produced language.

In order to answer the third research question about whether any existing differences are statistically significant, the Wilcoxon Signed-Rank Test was run to compare the different groups. This test was chosen due to the non-normality of distribution exhibited by the results obtained in this study, which makes the Wilcoxon test an appropriate choice since it does not assume normality of data distribution.

4. Results and Discussion

The present study aimed to determine whether there were differences in the use

of hedging words and engagement markers between essays generated by AI (obtained from ChatGPT-4o) and essays written by humans. In order to do this, two main collections of essays (*AI-Generated* & *Human-Written*) were compiled from four sub-collections (*AI-Default* & *AI-Humanized* and *Human-Hand-written* & *Human-Typed*). Each of these four groups of essays consisted of about 10,000 words, making up two main corpora of about 20,000 words and a total corpus of nearly 40,000 words, which is the data analyzed in this study. Overall raw frequencies for *hedges* and *engagement markers* instances are provided in Table 4.

Table 4. Hedging words and engagement markers across the four different sub-corpora as well as the two main corpora (AI Overall (1 & 2) and Human Overall (3 & 4)).

Corpus	Linguistic Feature	Hedging Devices		Engagement Markers	
		Tokens	Mean	Tokens	Mean
1	AI-Default	404	10.92	223	2.35
2	AI-Humanized	408	11.03	217	2.29
3	Human-Hand-Written	279	7.55	251	2.65
4	Human-Typed	221	5.97	284	2.99
5	AI-Generated (AI Overall)	405.5	11	220	2.32
6	Human-Written (Human Overall)	250	6.77	267.5	2.82

Note: The mean here is the outcome of dividing frequency counts by 37 (the number of target hedges) for hedging results, and by 95 (the number of target engagement markers) for engagement markers results.

According to Table 4, the total number of hedges found in the AIGEs is 405.5 (obtained by dividing the total number of the two AI sub-corpora *Default* (404) & *Humanized* (408) by 2), with a mean of 11 (obtained by dividing 405.5 by the total number of hedging words examined here, which is 37). Following the same method of calculation for the other two sub-corpora (*Human-Hand-written* & *Human-Typed*), the number of hedges found in the HWEs is 250, with a mean of 6.77.

Similarly, as shown in Table 4, the number of engagement markers obtained for the AIGEs (by adding the total numbers from the two AI sub-corpora and dividing the outcome by 2) is 220, with a mean of 2.32 (divided by the total number of engagement marker items examined here, which is 95). The number of engagement markers found in the HWEs is 267.5, with a mean of 2.82.

4.1. Hedging

In order to answer the first research question about the differences in hedging words between the two main groups of essays (i.e., *AI-generated* and *Human-written*), the findings presented in Table 4 and Figure 2 show that the AIGEs exhibited much more usage of hedging words (405.5) than did essays written by

humans (250). This was found to apply across all four sub-corpora. In other words, both of the AI essay sub-corpora showed higher instances of hedging words than both of the student essay sub-corpora. This large use of hedging found in the data is expected, considering that hedging is used extensively in academic discourse due to its important pragmatic functions (Hyland, 1996) and that it has recently been used much more frequently in academia than in the past (Hyland, 2011).

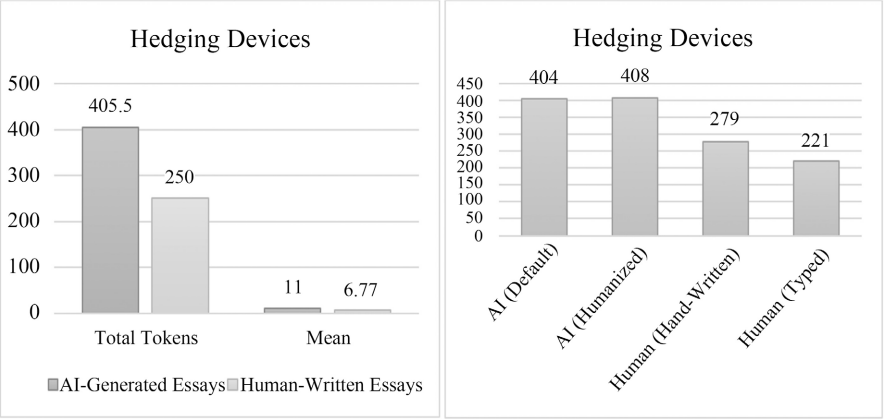

Figure 2. Hedging words in the four basic sub-corpora (right) from which the two main corpora (left) are compiled.

However, a question arises regarding why AIGEs show more usage of hedging than HWEs. Zhang and Crosthwaite (2025) noted that AI-produced language uses a more scientific and academic style (i.e., more abstract, conceptually-oriented, and less experientially-oriented). Thus, the present result could be understood in terms of AI-produced essays showing more academic features than human-written essays. In order to test the validity of this interpretation, lexical density results have been obtained from the UAM Corpus Tool. It was found that AIGEs showed more lexical density than HWEs (AI-Default: 60, AI-Humanized: 58, Human-Hand-written: 51, & Human-Typed: 49.50).

In order to answer the third research question about whether differences found between AIGEs and HWEs were statistically significant, a Wilcoxon rank sum test with continuity correction was run using the free statistical software RStudio. Since the data were not normally distributed, the non-parametric Wilcoxon test was appropriate to deal with such data. Although the descriptive statistics provided in **Figure 2** show there is a difference between the means, there was no statistical significance between AI essays and student essays in the use of hedging, W = 723.5, *p*-value ≥ 0.05. This finding could undermine the differences found between AI-produced language and human-produced language with respect to hedging as a linguistic feature. This non-significant result could also be because of the relatively small sample size (40,000 words), which may not be representative enough of the two types of texts or the linguistic features targeted here. As mentioned above, the case could simply be that there are differences in the use of hedging between AI

and Human writers, but they do not reach a distinguishing status between the two. This is supported by the difficulty faced by many journal article reviewers in distinguishing between abstracts produced by AI from those produced by humans, as reported by Casal and Kessler (2023).

If there is no statistical significance in the differences between AI and human essays with regard to the use of hedging, this means that the descriptive differences found here could be specific to this specific sample of texts and may not be similarly observable in other samples. It could also mean that the characteristics between AI-generated text and human-written text are, after all, not that distinguishable, which would have implications on the reliability of AI-text detection tools that some teachers/instructors use, as some researchers have previously highlighted (e.g., AlAfnan & MohdZuki, 2023; Bellini et al., 2024). This is critical because it may result in unfair judgments and penalties affecting students who submit their own work, but find out that these tools reported their work as produced by AI. On the other hand, with regard to concerns related to academic integrity, this might also have its drawbacks on ethical practices in some cases, such as the case where a student submits an AI-produced work as their own.

This calls for more research in an attempt to understand what distinguishing characteristics AI text may exhibit, relying on human expert judgments. While the overall AI's ability to detect content originality still does not seem ready to be relied upon (Chaka, 2023), there is also some evidence that when these tools show accuracy, they are much more accurate with AI-generated text than with human-produced text (Elkhatat et al., 2023). That is, most of the time, they are likely to give inaccurate judgments on human-produced texts, reporting that they are AI-generated. This is detrimental because it runs into the possibility of applying unfair penalties to students who are actually abiding by academic integrity guidelines.

A further qualitative look into the data has been taken to see what trends might be occurring. With regard to *modals*, the modal *can* was found to be the most frequently used modal in the two main corpora (341 times by AIGEs and 65 by HWEs). On the other hand, while the modal *may* was used 119 times in AIGEs, it was used only 13 times in HWEs. Conversely, while the modal *will* was used 55 times in HWEs, it was used only 15 times in the AIGEs. Out of the hedging *verbs* examined, while the most frequently used verb in AI essays was *suggest*, it was *argue* that was most commonly used in student essays. With regard to hedging *adjectives*, the most frequent adjective in AI and Student essays was *many*, with a frequency of 48 and 59, respectively. In a similar way, the most commonly used *adverb* by the two groups of essays was *about,* with a frequency of 51 (AI essays) and 60 (student essays). These findings, especially the one involving the modal *may*, are in line with what has been highlighted in the literature (e.g., Hyland, 1996) regarding the notion of objectivity in academic discourse, with which AI-produced language is more associated.

4.2. Engagement

With regard to the second research question about which group of essays showed

more use of engagement markers, **Table 4** and **Figure 3** show that HWEs exhibited more usage of engagement markers than did AIGEs, and this is found to be true across the four sub-corpora. This is in line with previous findings from Jiang and Hyland (2024b), where it was found that university student essays used more engagement markers than AI-produced essays. This finding could also be taken to corroborate Casal and Kessler's (2023) findings, where it was reported that journal reviewers were more certain that a research article's abstract was not produced by AI because of some instances of words or expressions that are more likely to be produced by human writers.

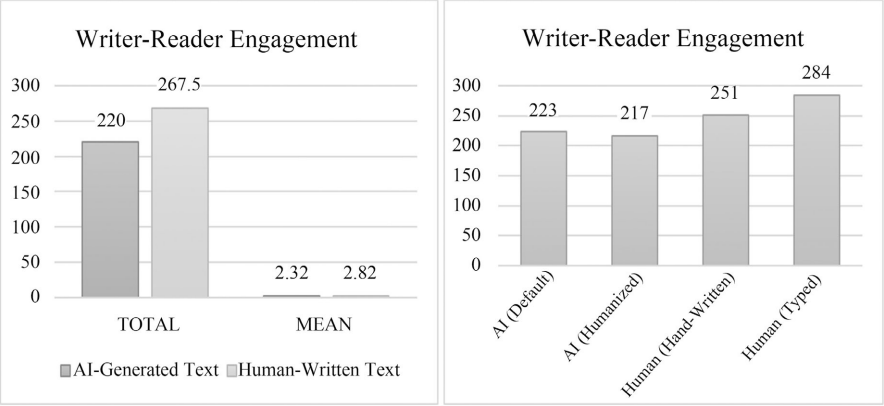

Figure 3. Engagement markers in the four basic sub-corpora (right), from which the two main corpora (left) are compiled.

As can be seen in **Figure 3**, there is a differing pattern of results. Whereas AIGEs used more hedging devices, HWEs show more usage of engagement markers. Engagement markers are categorized, according to Hyland (2005), into five categories (i.e., *reader mentions*, *questions*, *appeals to shared knowledge*, *directives, & personal asides*). As shown in **Figure 3**, the Human-written essay corpus (composed of the *Human-Hand-written* & *Human-Typed* sub-corpora) shows more usage of engagement markers (267.5) than AIGEs (220). As mentioned previously (when addressing hedging results), these numbers are obtained from adding the total numbers of Human-Hand-written (251) and Human-Typed (284), and dividing the outcome by two, which results in a mean value of 267.5. Similarly, the mean for the AIGEs (220) was obtained by adding the total numbers of *AI-Default* (223) and *AI-Humanized* (217), and dividing the outcome by two. The two mean values on the left side of **Figure 3** are obtained by adding the total numbers of the two basic corpora for each group (i.e., AI-generated & Human-written) and dividing the outcome by the number of items searched for as examples of engagement markers (95 in this case).

In order to see whether this difference was statistically significant, a Wilcoxon rank sum test with continuity correction was run using the free statistical software RStudio. Similar to the results of hedging word usage, the Wilcoxon test showed no statistical significance for the difference observed in the usage of engagement

markers between AI-generated and human-written essays $W = 4060$, p-value > 0.05. Similar implications to those mentioned earlier about the non-statistically significant result of "hedging" apply here as well. To be more specific, these implications are directly related to hedging and engagement markers as "distinctive features" of AI-generated and human-written texts and not about "hedging" and "engagement markers" as discourse markers.

A deeper look at some examples of engagement marker usage by human writers and AI (ChatGPT) may give a clearer idea of what differences might be occurring here. For the category *reader mention*, for example, the personal plural pronouns *we* and *our* were found in student essays 23 and 38 times, respectively, whereas they were used by ChatGPT 13 and 11 times, respectively. Within the HWEs, these pronouns were used much more by the Human-Typed (*more recent/online*) essay group (21 & 24 times) compared to the Human-Hand-written (12 & 14 times). This may suggest that there is an increasing trend in using engagement markers in academic writing by human writers. Below are some examples of this usage across the two main corpora (*AI-generated* and *Human-written*):

1. *Why should* ***we*** *be concerned with the life of a violent criminal?* (Student Essay)
2. *… it is usually based upon the rage of* ***our*** *society towards a criminal's violent act*. (Student Essay)
3. *… it's essential to define what* ***we*** *mean by creativity in the educational context*. (AI Essay)
4. *… technology can serve as a means of connection, especially in* ***our*** *increasingly globalized world*. (AI Essay)

It is apparent from Examples 1 and 2 above that the writer uses the pronouns *we* and *our* for the purpose of engaging the reader as being part of a larger group (i.e., society). Employing this engaging strategy, the writer seems to be trying to convince the reader that, since they both belong to the same group, what they think is best for them is also best for the reader. This is also a clear example of a characteristic mentioned in the literature (e.g., Zhang & Crosthwaite, 2025) specific to Human-writer essays that are related to the real-world experience writers have. In Example 1, the writer is referring to a controversial issue taking place in society, and Example 2 shows "society" as an entity of belonging. On the other hand, while Examples 3 and 4 use the same pronouns (we & our), they are used in more abstract contexts.

This reader-engaging strategy is clearly depicted by Hyland's (2005) model through the category "reader mentions/pronouns". These linguistic items can largely be used by writers to engage readers by showing how similar, how close, and sometimes how much the writer's experiences and feelings resemble what the reader experiences and feels. Writers might use "we" to make the reader feel that they belong to the same group or share the same identity as the writer, and they may use "our" to go beyond relational aspects to show that they are very close, to the degree that they share a possession of some kind. These subtle—but deeply del-

icate—meanings are much more vivid in examples 1 and 2 than in examples 3 and 4. That is, the usage of these pronouns in examples 1 and 2 is more alive and moving than in 3 and 4, which seems dry because of abstractness. This emphasizes the significant role the social context plays in distinguishing human-written text, as Zhang and Crosthwaite (2025) have pointed out.

With regard to questions as engagement markers, HWEs included 22 questions, whereas AIGEs included only four. The first question below was by a student, and the second was by AI:

1. *What does this have to do with the authors argument?* (Student Essay)

2. *Should they encourage their young aspirations, or should they draw a line in the sand regarding safety?* (AI Essay)

It could be argued that the first question above is an open-ended question that allows the reader to freely think about possible responses, which encourages more engagement with what the writer is saying. The second question, on the other hand, is a closed-ended question restricting the reader's options to only two possibilities, neither of which may be accepted by the reader.

With regard to directive verbs, where the writer typically engages the reader to take some mental or material action, the verb need was frequently used in HWEs (14 instances), and consider was frequently used in AIGEs (24 instances). The following are examples of the use of need and consider in the corpus:

1. *Proponents **need** to argue that the effects of discrimination without quotas would be worse than the minor failures that occur because they are expected to.* (Student Essay)

2. ***Consider** the world of sports, where raw talent can get an athlete noticed, but it is sustained effort that determines long-term success.* (AI Essay)

4.3. Limitations of the Study

One main limitation of the present study is that, whereas inferential statistics were used to find out if any occurring differences were most likely due to chance, the size of each corpus is relatively small to make overall generalizations about the four types of text represented by the four sub-corpora. Therefore, while conclusions may be drawn for the specific corpora used here, more research with larger corpus data is needed before generalizable conclusions can be made. Even though the findings are not entirely novel (i.e., similar findings have been separately reported in previous studies), considering this is a mainly quantitative study (with a small-sized corpus), it would be an overstatement to conclude that hedging and engagement are distinctive features of AI-generated text and human-produced text, respectively. While these features could be broadly taken as indicative of AI-generated and human-produced texts, larger corpus data and more qualitative analysis would ensure clear insights into these differences and how distinctive of each type of text they may be.

5. Conclusion

The present study aimed to find out whether there were differences based on which

AI-generated text and human writing can be distinguished. Two corpora (20,000 words each) were analyzed and compared based on the use of two main characteristics of academic writing (hedging and engagement). Descriptive statistics showed noticeable differences in which AI-generated text was associated more with the use of hedging and human writing was associated more with the use of engagement markers. However, these differences were not found to be statistically significant.

For the time being, while AI is capable of producing advanced academic texts, as reflected by the high lexical density and the use of hedging devices, which is a relatively recent feature of academic writing, it still seems to struggle with making use of other features exhibited by academics, such as engagement markers, which build on and rely on personal experiences and social contexts. Overall, according to what has been reported in the literature and based on these findings, it could be said that differences between AI-generated texts and human-produced texts exist, although pinpointing the exact nature of these differences still calls for more research.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Adrian, D., & Fajri, M. S. A. (2023). Hedging Practices in Soft Science Research Articles: A Corpus-Based Analysis of Indonesian Authors. *Cogent Arts & Humanities, 10,* Article ID: 2249630. https://doi.org/10.1080/23311983.2023.2249630

AlAfnan, M. A., & MohdZuki, S. F. (2023). Do Artificial Intelligence Chatbots Have a Writing Style? An Investigation into the Stylistic Features of ChatGPT-4. *Journal of Artificial Intelligence and Technology, 3,* 85-94. https://doi.org/10.37965/jait.2023.0267

Alshalan, K. F., & Alyousef, H. S. (2024). A Corpus-Based Study of the Experiential Meaning in Business Students' Hand-Written and ChatGPT-Generated Argumentative Essays. *Middle East Research Journal of Linguistics and Literature, 4,* 93-103. https://doi.org/10.36348/merjll.2024.v04i05.002

Anthony, L. (2022). *AntConc (Version 4.3.1) [Computer Software].* https://www.laurenceanthony.net/software

Bellini, V., Semeraro, F., Montomoli, J., Cascella, M., & Bignami, E. (2024). Between Human and AI: Assessing the Reliability of AI Text Detection Tools. *Current Medical Research and Opinion, 40,* 353-358. https://doi.org/10.1080/03007995.2024.2310086

Casal, J. E., & Kessler, M. (2023). Can Linguists Distinguish between ChatGPT/AI and Human Writing? A Study of Research Ethics and Academic Publishing. *Research Methods in Applied Linguistics, 2,* Article ID: 100068. https://doi.org/10.1016/j.rmal.2023.100068

Chaka, C. (2023). Detecting AI Content in Responses Generated by ChatGPT, YouChat, and Chatsonic: The Case of Five AI Content Detection Tools. *Journal of Applied Learning and Teaching, 6,* 94-104.

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the Efficacy of AI Content Detection Tools in Differentiating between Human and AI-Generated Text. *International Journal for Educational Integrity, 19,* Article No. 17.

https://doi.org/10.1007/s40979-023-00140-5

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2024). A SWOT Analysis of ChatGPT: Implications for Educational Practice and Research. *Innovations in Education and Teaching International, 61,* 460-474. https://doi.org/10.1080/14703297.2023.2195846

Hyland, K. (1996). Writing without Conviction? Hedging in Science Research Articles. *Applied Linguistics, 17,* 433-454. https://doi.org/10.1093/applin/17.4.433

Hyland, K. (2002). Options of Identity in Academic Writing. *ELT Journal, 56,* 351-358. https://doi.org/10.1093/elt/56.4.351

Hyland, K. (2005). Stance and Engagement: A Model of Interaction in Academic Discourse. *Discourse Studies, 7,* 173-192. https://doi.org/10.1177/1461445605050365

Hyland, K. (2011). Academic Discourse. In K. Hyland, & B. Paltridge (Eds.), *Continuum Companion to Discourse Analysis* (1st Ed.). Bloomsbury Publishing.

Hyland, K., & Jiang, F. (2017). Is Academic Writing Becoming More Informal? *English for Specific Purposes, 45,* 40-51. https://doi.org/10.1016/j.esp.2016.09.001

Jiang, F., & Hyland, K. (2024a). Does ChatGPT Argue Like Students? Bundles in Argumentative Essays. *Applied Linguistics, 46,* 375-391. https://doi.org/10.1093/applin/amae052

Jiang, F., & Hyland, K. (2024b). Does ChatGPT Write Like a Student? Engagement Markers in Argumentative Essays. *Written Communication, 42,* 463-492. https://doi.org/10.1177/07410883251328311

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F. et al. (2023). ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences, 103,* Article ID: 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kim, M., & Adlof, L. (2024). Adapting to the Future: ChatGPT as a Means for Supporting Constructivist Learning Environments. *TechTrends, 68,* 37-46. https://doi.org/10.1007/s11528-023-00899-x

Lu, Q., Yao, Y., Xiao, L., Yuan, M., Wang, J., & Zhu, X. (2024). Can ChatGPT Effectively Complement Teacher Assessment of Undergraduate Students' Academic Writing? *Assessment & Evaluation in Higher Education, 49,* 616-633. https://doi.org/10.1080/02602938.2024.2301722

Mizumoto, A., & Eguchi, M. (2023). Exploring the Potential of Using an AI Language Model for Automated Essay Scoring. *Research Methods in Applied Linguistics, 2,* Article ID: 100050. https://doi.org/10.1016/j.rmal.2023.100050

O'Donnell, M. (2018). *UAM Corpus Tool (Version 6.2) [Computer Software].* http://www.corpustool.com

OpenAI (2023). *ChatGPT: Optimizing Language Models for Dialogue.* https://openai.com/blog/chatgpt

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic Review of Research on Artificial Intelligence Applications in Higher Education—Where Are the Educators? *International Journal of Educational Technology in Higher Education, 16,* Article No. 39. https://doi.org/10.1186/s41239-019-0171-0

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M. et al. (2021). A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity, 2021,* Article ID: 8812542. https://doi.org/10.1155/2021/8812542

Zhang, M., & Crosthwaite, P. (2025). More Human than Human? Differences in Lexis and Collocation within Academic Essays Produced by ChatGPT-3.5 and Human L2 Writers. *International Review of Applied Linguistics in Language Teaching*, 1-28.

https://doi.org/10.1515/iral-2024-0196

Zhou, X., Gao, Y., & Lu, X. (2023). Lexical Complexity Changes in 100 Years' Academic Writing: Evidence from Nature Biology Letters. *Journal of English for Academic Purposes, 64,* Article ID: 101262. https://doi.org/10.1016/j.jeap.2023.101262