

Engagement Functions of AI-Generated Text from the Perspective of Appraisal Theory

Sun Hechu

Nanjing University of Aeronautics and Astronautics, Nanjing, China, 211106

Received: 22 Aug 2026; Received in revised form: 19 Sep 2026; Accepted: 24 Sep 2026; Available online: 28 Sep 2026

©2026 The Author(s). Published by Infogain Publication. This is an open-access article under the CC BY license

(<https://creativecommons.org/licenses/by/4.0/>).

Abstract— Within the framework of Appraisal Theory, this corpus-based study investigates the use and functions of engagement strategies in human-written and AI-generated academic discourse. A total of 190 academic texts (95 human-written from the BAWE corpus and 95 generated by DeepSeek-V3) were analyzed. A mixed-methods approach was adopted, involving both quantitative and qualitative analyses of Dialogic Expansion and Dialogic Contraction strategies. The results show that human writing contains significantly more engagement resources than AI-generated texts overall. This difference holds for both Dialogic Expansion and Dialogic Contraction. AI-generated texts also exhibit frequent vague Attribution and localized overuse of Proclaim resources. These patterns are associated with weaker stance construction, less effective dialogic negotiation, and weaker reader-oriented persuasion. Based on the observed differences, this study offers implications for academic writing instruction and human-AI collaboration.



Keywords— Appraisal Theory, Engagement system, AI-generated text, academic writing, corpus-based study

I. INTRODUCTION

Large Language Models (LLMs) have advanced rapidly in recent years, demonstrating substantially enhanced capabilities in language-intensive tasks such as translation, literary creation, and text polishing. Yet questions remain about their interpersonal and dialogic meaning-making in academic discourse. Even when state-of-the-art large language models produce grammatically fluent and logically coherent texts, readers can still discern subtle differences between AI-generated text and human writing. The measurement of such differences takes two principal experimental approaches: human judgment and machine detection. Human-based methods typically present AI-generated and human-written texts to raters of varying academic proficiency for identification or scoring. Findings

are markedly divided. Some scholars argue that the boundary between AI-generated and human language is increasingly blurred, and that AI can achieve levels indistinguishable from human writing (Köbis & Mossink, 2021). Even professional reviewers find it nontrivial to distinguish between the two (Casal & Kessler, 2023). Other scholars contend that AI writing actually surpasses human performance in academic contexts, yielding higher overall scores and overwhelming advantages in lexical richness, syntactic complexity, discourse cohesive devices, and grammatical error rate (Herbold et al., 2023). Still others maintain that AI texts are easily identifiable, as they exhibit serious defects both at the surface linguistic level (Turós et al., 2026; Qiu & Pan, 2026) and in pragmatic effects (Pütz & Esposito, 2024). This Turing-test-style “authenticity

judgment” is participants’ subjective judgments. It does not focus on measurable linguistic features but relies heavily on subjective perceptions.

Machine detection, by contrast, employs tools to identify AI-generated texts through methods such as traditional machine-learning classifiers, BERT/RobERTa fine-tuning, and perplexity-based detectors like GPTZero, achieving identification accuracy close to 99% (Berber Sardinha, 2024; Barrot & Aranda, 2025). Compared with human detection, these methods can detect AI-involved writing very precisely. Such studies have found that AI-generated texts feature higher frequencies of special symbol usage, greater lexical complexity, and longer and more complex sentences, but lack rhetorical and emotional expression (Zhang & Crosthwaite, 2026). These approaches do not interpret textual meaning; they cannot unpack the underlying intellectual substance of a text. In essence, they extract surface-level linguistic features from the text to classify its category. The linguistic difference between AI and human writing thus becomes a detection criterion, yet the underlying reasons for the difference remain unexplained. In other words, these detection methods answer the question “Was this written by AI?” rather than “Why does AI writing not read like human writing?” Addressing the latter question requires an analytical tool that can penetrate the linguistic surface and reach the communicative functions beneath.

From the perspective of Systemic Functional Linguistics, language fulfills three metafunctions: ideational, interpersonal, and textual (Halliday, 1985). AI texts generally convey propositional content fluently, although their factual accuracy may vary. However, AI-generated texts still exhibit deficiencies in interpersonal meaning-making, where authorial stance is required. Existing research has mostly focused on surface-level linguistic features such as lexical complexity, part-of-speech patterns, collocation, and sentence length, only a small number of studies have examined AI texts from the perspective of Appraisal Theory.

Against this backdrop, the present study draws on Martin and White’s Appraisal Theory framework (Martin & White, 2007) to compare the use of engagement resources between AI-generated and human-written texts, with particular attention to whether these resources perform their

communicative functions at specific discourse locations. The following research questions are addressed:

- (1) What differences exist between AI-generated and human-written academic texts in terms of the realization of engagement?
- (2) What patterns do these differences exhibit?
- (3) What implications can be drawn from these patterns?

II. THEORETICAL FRAMEWORK

Appraisal Theory represents a development in the study of interpersonal meaning within Systemic Functional Linguistics. The engagement system is concerned with the dialogic relationships among different voices in discourse. Through dialogic expansion and dialogic contraction, it regulates the dialogic space between the author’s stance and other voices (Martin & White, 2007). Dialogic expansion and contraction directly address how writers acknowledge or suppress alternative positions, how they leave dialogic room for readers and other voices, and how they construct the multi-voiced negotiability that academic discourse should possess—which corresponds precisely to the interpersonal metafunction in Systemic Functional Linguistics.

Writers deploy engagement resources to express their own stance and, as members of a particular community, build a communicative bridge with readers. Texts that rely predominantly on monoglossic assertions with limited engagement resources tend to restrict heteroglossic dialogic space, and may fail to realize the negotiability expected in much academic argumentation. Monoglossic statements are themselves a legitimate rhetorical option in academic writing; the concern in AI-generated texts lies in their over-reliance on monogloss with insufficient heteroglossic negotiation. From the perspective of the Appraisal Theory, the most salient difference is that human-written texts typically carry richer interpersonal meaning, especially in terms of engagement resources. Compared with AI texts, humans use more diverse multi-voiced resources, for example, entertain and acknowledge, with a more balanced distribution of dialogic expansion and contraction, and with stronger reader-oriented stance expression.

Within the engagement system, which centers on heteroglossic voices, Dialogic Contraction strategies serve to close down or constrain dialogic space, while Dialogic

Expansion strategies open up space for alternative voices. contraction includes disclaim (deny and counter) and proclaim (pronounce, concur, and endorse), whereas Expansion includes entertain and attribute (acknowledge and distance), see Fig.1. This framework provides the analytical tool for examining how writers position themselves vis-à-vis alternative positions and how they negotiate the dialogic space of academic argumentation.

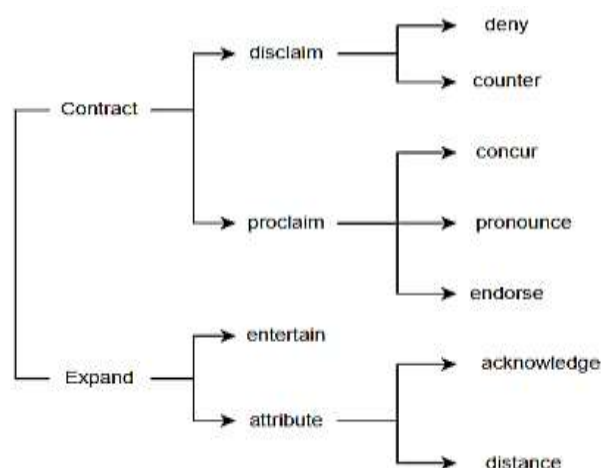


Fig.1: Engagement System

III. LITERATURE REVIEW

Scholars have explored engagement in academic writing primarily from two theoretical perspectives. The first is Martin and White's (2007) Appraisal Theory model, and the second is Hyland's (2005) interpersonal interaction model. Existing research has drawn on both models to investigate how academic writers position themselves in relation to readers, texts, and the academic discourse community. For instance, Jiang and Hyland (2025a), taking metadiscourse nouns as their analytical point of entry, found that ChatGPT prefers to use way nouns to describe research methods, processes, or forms, while students use state nouns for evaluative reasoning and evidential nouns to establish empirical foundations. This suggests that AI tends to describe "what is" rather than evaluate "how it is." Subsequently, Jiang and Hyland (2025b) compared ChatGPT-generated argumentative essays with those written by British university students and found that ChatGPT texts performed well in structural coherence and logical organization, but exhibited significantly lower frequencies of interactional metadiscourse, with marked reductions in hedges, boosters, and attitude markers. As a

result, AI texts presented an impersonal expository tone, whereas student texts demonstrated stronger rhetorical engagement and reader interaction awareness.

A study of academic book reviews similarly confirmed that ChatGPT overall uses interactive metadiscourse more frequently, but this frequency stems from the overuse of attitude markers, while hedges and self-mentions are significantly insufficient. AI texts thus rely on overt evaluative language but lack nuanced deliberation and authorial presence (Yao & Liu, 2025). Cross-disciplinary abstract research has likewise noted that AI's weakness in interactive metadiscourse manifests as "limited stance modulation and reduced personal involvement," coupled with insufficient sensitivity to disciplinary rhetorical norms (Zhang & Zhang, 2025). These studies consistently point to the rhetorical shortcomings of AI texts: a bias toward expository expression, weakened authorial stance, and inadequate deployment of dialogic engagement strategies. Although the interpersonal interaction model provides an effective analytical tool for examining authorial voice construction and reader engagement in academic texts, Martin and White's (2007) Appraisal Theory framework offers a more comprehensive account of engagement phenomena. It encompasses a rich repertoire of strategies and sub-strategies that can more fully interpret the author-reader interaction. Engagement is a key component of discourse construction: writers rely on engagement to conduct dialogic negotiation, establish credibility, and shape persuasive tone, thereby accomplishing meaning construction. Engagement research within Appraisal Theory examines how writers deploy linguistic resources to realize dialogic expansion and dialogic contraction, so as to participate in the academic discourse community.

Among existing studies on AI-generated texts, those specifically investigating engagement resources remain scarce. The present study compares engagement markers in human-written and AI-generated academic argumentative essays, focusing on Dialogic Expansion and Dialogic Contraction strategies.

IV. METHODOLOGY

4.1. Data Collection

This study examines whether the engagement and appraisal linguistic features invoked by AI-generated texts can serve

the interpersonal and persuasive communicative functions of the target texts.

First, a parallel comparable corpus was constructed. Human-written texts were drawn from the British Academic Written English (BAWE) corpus. A stratified sampling method was used to select 95 academic texts, covering 20 disciplinary themes and encompassing genres such as argumentative essays, critical reviews, expository texts, case studies, and research reports. The disciplinary range spans humanities, social sciences, natural sciences, medicine, and various applied sciences. Considering that large language models have been extensively used for academic text assistance since 2022, and to exclude interference from AI rewriting or polishing, all human-written texts were completed before 2022, thereby constructing a sub-corpus of pure human writing free from AI-assisted elements.

The DeepSeek-V3 model was selected to generate the AI sub-corpus. This model possesses strong language processing capabilities and can effectively capture internal logical relations and semantic features, enabling high-quality text generation under complex tasks and ensuring the stability and reliability of the corpus output. The AI sub-corpus follows the principle of parallel corpus construction, matching the human corpus one-to-one in terms of disciplinary theme, genre type, text number, and length. Standardized prompts were designed; taking a linguistics text as an example, the prompt is shown as below:

You are a student majoring in Linguistics (Discipline). Please compose an essay (Text-type) on the topic of "Chimpanzees can use signs but do they have language?" (Topic), with a word-count requirement of 500 (Word Count).

4.2. Data Coding

After the two sub-corpora were constructed, the UAM Corpus Tool was used as the annotation tool. Drawing on the engagement system within Martin and White's (2007) Appraisal Theory, lexical and grammatical forms of various engagement resources were retrieved and annotated, forming an initial engagement resource marker list. The engagement system focuses on dialogic relationships among different voices in discourse, primarily including Dialogic Expansion and Dialogic Contraction. To avoid misclassification based solely on surface forms, the

annotation process was guided by Martin and White's (2007) distinction between engagement resources and propositional content, with each case judged individually in its specific discourse context.

Ten texts were randomly sampled from the two parallel corpora as a reliability-check sub-sample. Two graduate students in linguistics conducted independent annotation. The Cohen's Kappa coefficient was calculated as $k = 0.76$, indicating a high level of coding agreement.

4.3. Data Analysis

Following formal annotation, the study conducted a functional validity analysis of every engagement marker in both corpora. Each marker was examined for whether its claimed function was actually realized in context, thereby exploring how effectively the engagement resources generated by large language models function in actual academic communication.

V. RESULTS

Table 1 presents the basic information of the two corpora constructed in this study. Table 2 shows the raw and normalized frequencies of engagement resources across the two corpora. In terms of frequency data, the normalized frequency of engagement markers in human writing is 21.98 per 1000 words, compared with 12.69 in AI-generated texts. The difference is highly significant ($LL = 191.02$, $p < 0.001$), indicating that AI-generated texts deploy engagement resources considerably less frequently overall.

Table 1: Corpus Information

Corpus	Number of texts	Word count
Human-written	95	74016
AI-generated	95	74998

When broken down by subsystems, Dialogic Contraction resources in AI texts amount to only 3.97 per 1000 words, compared with 9.25 in human writing ($LL = 162.60$, $p < 0.001$), while Dialogic Expansion resources also occur significantly less frequently in AI texts (8.72 vs. 12.73 per 1000 words, $LL = 56.94$, $p < 0.001$). Thus, AI-generated texts are deficient in both dialogic expansion and contraction resources, with the shortage of contraction resources being particularly pronounced: AI contraction resources amount to only 42.9% of those in human writing, whereas expansion resources reach 68.5% of the human level.

Table 2: Frequencies of Engagement Resources

	Human raw count	Human normalized (/1000 words)	AI raw count	AI normalized (/1000 words)	LL	p
Engagement	1627	21.98	952	12.69	191.02	$p < 0.001$
Expansion	942	12.73	654	8.72	56.94	$p < 0.001$
Contraction	685	9.25	298	3.97	162.60	$p < 0.001$

VI. DISCUSSION

However, frequency differences alone cannot reveal functional differences. The core function of engagement resources lies not in their formal occurrence, but in whether they are grounded in a genuine anticipation of readers' knowledge states, reading difficulties, and potential questions, and whether they perform a real dialogic function at specific discourse positions.

6.1. Weakened Stance Construction

In academic discourse, the primary function of engagement resources is to help writers construct a recognizable and locatable academic stance. Whether through Attribution resources that introduce external voices, self-mention that marks the writer's research actions and judgments, or Dialogic Contraction resources that express degrees of certainty, the fundamental purpose is to enable readers to perceive "where the author stands." However, AI-generated texts exhibit a systematic weakening of this function, consistent with the overall scarcity of engagement resources observed in the corpus data (12.69 vs. 21.98, $p < 0.001$).

The most prominent manifestation is the extreme scarcity of first-person engagement. First-person engagement is the core means by which authors explicitly refer to themselves in discourse and project the self as an academic subject. Human writers use "I conducted semi-structured interviews" to mark research actions (Example 1), use "I don't believe" to express personal judgment (Example 2), and use "we can see that" to invite readers into shared reasoning (Example 3). Together, these resources construct an academic subject with a stance, a judgment, and dialogic awareness. AI, by contrast, almost entirely avoids first person, preferring "It is believed that" over "I believe that," and dissolving authorial presence through passive voice and impersonal structures. From the perspective of the engagement system, "It is believed that" is an acknowledge

resource that attributes the proposition to a vague collective voice; while formally introducing an external voice, it substantively dissolves the author's own stance expression.

(1) I conducted semi-structured interviews with...

(2) Whilst there have been famous cases of chimps being taught to produce signs, I don't believe it can be said that chimpanzees have language.

(3) When looking at these design features with reference to chimpanzees, we can see that the sign language they are using does indeed confer to the rules of duality of patterning.

The inadequacy of attribution resources further weakens stance construction. Human writers typically specify the author, year, and content of the cited view, thereby introducing a genuinely identifiable dialogic partner into the discourse. AI, by contrast, frequently relies on vague attributions such as "Some scholars contend that" or "Research has shown that" (Examples 4 and 5). While such expressions formally introduce an external voice, they often fail to establish a substantive dialogic relationship with the actual academic community. Consequently, the academic foundation of the author's stance becomes less secure.

(4) For decades, primatological research has confirmed that...

(5) Some scholars contend that...

6.2. Illusory Multi-voiced Dialogic Negotiation

Academic discourse is essentially a site of multi-voiced negotiation. When advancing knowledge claims, writers need to acknowledge alternative positions and establish the legitimacy of their own views through dialogic engagement. Within the engagement system, Dialogic Expansion and Dialogic Contraction resources are precisely the linguistic mechanisms that realize such multi-voiced negotiation. In AI-generated texts, however, this function is markedly weakened.

The first manifestation is the overall scarcity of dialogic negotiation resources. As the frequency data show, AI texts deploy fewer engagement resources across the board (12.69 vs. 21.98, $LL = 191.02$, $p < 0.001$), indicating that AI-generated academic discourse is comparatively monologic. Within this overall scarcity, Dialogic Contraction resources are disproportionately depleted: AI contraction resources amount to only 42.9% of those in human writing (3.97 vs. 9.25, $LL = 162.60$, $p < 0.001$), while Dialogic Expansion resources reach 68.5% of the human level (8.72 vs. 12.73, $LL = 56.94$, $p < 0.001$). This pattern reveals that the multi-voiced negotiation of AI texts is doubly deficient: it lacks both the dialogic space opening provided by expansion resources and the dialogic space management provided by contraction resources. Yet the contraction deficit is more severe, meaning that AI writers are especially poor at deploying Proclaim and Disclaim resources to negotiate the boundaries of dialogic space.

A second manifestation is the functional imbalance within contraction resources. Although AI uses fewer contraction resources overall, the proclaim resources it does deploy are often misdirected. As Example (6) illustrates, the proclaim resource “remarkably profound” assigns an inappropriately high evaluative intensity to an ordinary experimental limitation, constituting clear overstatement.

(6) This experimental limitation brings remarkably profound influences to the research paradigm.

The force of the tone is grossly disproportionate to the importance of the proposition itself, failing to reflect the actual argumentative context. From the engagement system perspective, “remarkably profound,” as a proclaim strategy within Dialogic Contraction, should function “to strengthen the certainty of a proposition on the basis of sufficient evidence.” However, the proposition here merely describes an experimental limitation and does not possess the argumentative necessity that warrants such strengthening. This excessive contraction produces an unnecessary compression of dialogic space, preventing readers from perceiving the author’s prudent assessment of the proposition’s certainty. Moreover, the entertain and proclaim resources triggered by AI result in convoluted sentences with unclear focus, creating “tone noise” that interferes with readers’ information reception.

The “pseudo-dialogue” of Attribution resources further hollows out multi-voiced negotiation. Examples (4) and (5) retain the formal attribution frame but provide no specific source, failing to fulfill the dialogic function that acknowledge resources are supposed to perform in the engagement system. Such pseudo-attribution creates an illusion of multi-voicedness without actually introducing an identifiable, answerable external voice; the dialogic negotiation thus loses its real adversarial counterpart. Such an acknowledge resource not only traces the information source but also introduces a genuinely identifiable dialogic partner into the discourse, laying the groundwork for subsequent negotiation, counterargument, or extended discussion.

The absence of coordination between first-person engagement and multi-voiced negotiation resources leaves multi-voiced negotiation without a subject to sustain it. Human writers use “I don’t believe” to refute while acknowledging the opposing view, demonstrating the coordinated operation of Dialogic Contraction and Expansion. AI texts as a whole present monologic contraction: the dialogic space is occupied by a single voice, and multi-voiced negotiation exists in name only.

6.3. Failed Reader-oriented Rhetorical Persuasion

The ultimate orientation of engagement resources is the reader. Academic writing is a special form of dialogue in which authors, through engagement resources, anticipate readers’ knowledge states, reading difficulties, and potential questions, and accordingly select rhetorical strategies to achieve persuasion. In Martin’s terms, this anticipatory orientation can be captured through the notion of the putative reader—the reader position that a text construes for itself through its deployment of engagement resources. AI-generated texts exhibit a clear failure in this function.

The reduced overall frequency of Entertain resources in AI texts (8.72 vs. 12.73, $p < 0.001$) already signals a weakened capacity to acknowledge readers’ possible alternative positions. Yet functional examination reveals that even the entertain resources that do appear in AI texts are not grounded in genuine reader anticipation. Human writers’ use of Entertain resources is grounded in anticipation of readers’ psychological states—adding hedges where readers may be skeptical, inserting reminders where

attention may wander, and deploying rhetorical questions where critical thinking is stimulated. AI's entertain resources are more a probabilistic matching of high-frequency modal and evidential patterns from training data than context-sensitive communicative behavior. The resources appear, but reader awareness is not genuinely present.

The absence of rhetorical questions is a typical manifestation of this persuasion failure. In Example (7), the human author constructs a multi-voiced negotiation dialogic space through consecutive questions: "Then what are the effects...?" prompts reader reflection, "Is it really so beneficial..." challenges an existing stance, and "Or to what extent..." further opens up the possibility of alternative positions. These rhetorical questions, as entertain strategies within Dialogic Expansion, acknowledge the existence of alternative positions and guide readers into evaluating those positions through questioning.

(7) Then what are the effects of transformational leadership in practice? Is it really so beneficial as 'indispensable' for organizational effectiveness? Or to what extent can it be viewed as 'unnecessary' and even potentially 'a dangerous curse'?

By contrast, AI writing more often adopts impersonal objective statements and avoids rhetorical questions. AI is monologic in essence: it has no genuine dialogic interlocutor. This means that although AI texts may be information-dense, they lack the warmth of genuine exchange and the spark of intellectual collision.

The failure of persuasion is also evident in the disconnection between Dialogic Contraction and reader anticipation. Proper Dialogic Contraction should be grounded in the anticipation of potential reader objections—where writers anticipate that readers will raise objections, they respond with evidence and argument, thereby accomplishing persuasion. AI's proclaim resource usage is not grounded in such anticipation but in probability-driven over-assertion. When AI does deploy contraction resources, it tends to use them where they are unwarranted, while failing to deploy them where they would genuinely persuade, that is, at points where readers are likely to object.

VII. CONCLUSION

Regarding the first research question, the study finds that AI-generated academic texts deploy engagement resources significantly less frequently overall than human-written texts, and this scarcity cuts across both Dialogic Expansion and Dialogic Contraction, with the contraction deficit being more pronounced. On the positive side, AI can construct well-layered textual structures and convey information with a surface neutrality that readers may find efficient. However, these advantages operate only at the linguistic formal level and still require human verification. In terms of actual communicative function, the deployment of AI engagement resources is not consistently effective and frequently reveals problems such as hollowed-out acknowledge, misdirected proclaim, formalistic entertain, and deficient first-person engagement. The reason lies in the fact that AI has not yet acquired human theory-of-mind reasoning and academic subjectivity; relying solely on statistical probabilities to replicate engagement resource templates cannot establish a genuine dialogic connection between engagement resources and discourse content, nor can it fully reproduce the speculative core and communicative value embodied in human writing.

In terms of the second research question, AI engagement resources generally exhibit the characteristic of formally replicating human usage while functioning within limited bounds, forming a distinctive style of AI discourse—namely, overall underuse of engagement resources, disproportionate contraction deficit, vague attribution, misdirected Proclaim, low person reference, and weak multi-voicedness. AI language need not be regarded as a clumsy copy of human language; it may gradually develop its own independent style. When humans use it, they should treat it as a resource rather than a substitute, and must not abandon verification and judgment merely because the output is fluent and neat.

For the third research question, the differences between AI and human engagement resource usage offer implications for student academic writing and AI-assisted writing. First, for student academic writing, the use of engagement resources should not mechanically replicate template expressions but should be deployed in context as needed. Students should be aware that both expansion and contraction resources are essential for building a credible

academic stance and negotiating with readers; an overall scarcity of engagement resources results in monologic, impersonal prose. Second, in AI-assisted academic writing scenarios, the human-machine collaboration model requires users to develop discriminatory awareness. One should leverage AI's efficiency advantages in text structuring and preliminary information integration, but must remain vigilant regarding the information AI outputs, especially its tendency to underuse dialogic negotiation resources and to deploy Proclaim where it is unwarranted. AI can provide scaffolding for text structuring and language polishing, while humans undertake the core tasks of academic judgment, logical reasoning, stance construction, and reader-oriented communication. AI excels at formal-level text organization; humans excel at context-sensitive speculation and interpersonal dialogue. Each has its strengths; the relationship is not one of simple substitution but of complementary collaboration.

From a human-centered perspective, people aspire to anthropomorphize artificial intelligence. They hope that the internal mechanisms and representational forms of models will emerge toward a "human-like mind," pursuing more human-like experiential expression and perceptual capacities. However, artificial intelligence currently cannot fully simulate the human mind, and therefore a certain gap remains between its linguistic expression and human expression. We cannot directly equate a higher degree of "anthropomorphism" in AI language with good language quality. Even humans cannot serve as a single gold standard. The present study's examination from the engagement perspective is primarily intended to reveal differences rather than to judge good or bad. As technology continues to develop, AI's human-like cognition, emotional expression, and reader awareness may strengthen, and generating texts indistinguishable from human writing is not impossible. The closer that day comes, the more humans must maintain critical scrutiny—benefiting from its efficiency while safeguarding the responsibility, judgment, and dialogicity inherent in language.

Several limitations should be acknowledged. First, AI texts were generated solely by DeepSeek-V3 with standardized prompts, so findings may not generalize to other LLM architectures or prompting regimes. Second, human texts were sampled from the pre-2022 BAWE corpus, restricting

diversity in writer backgrounds and genres. Third, annotation focused only on explicit lexical-grammatical engagement realizations; implicit engagement and overlaps between the Engagement and Graduation systems were left unexamined. Fourth, functional validity evaluations are interpretive. Although inter-annotator agreement for formal marker annotation was satisfactory, this measure derives from a relatively small subsample and does not apply separately to functional validity coding.

Future research could extend this work by adopting multiple LLMs and varied prompting regimes to test the generalizability of engagement-related patterns. It would also be fruitful to incorporate human blind-rating experiments to link textual linguistic features to readers' real-world perceptions.

REFERENCES

- [1] Barrot, J. S., & Aranda, Ma. R. R. (2025). Efficacy of AI-Text Detection Tools in Distinguishing Student-Produced, AI-Edited, and AI-Generated Essays. *Technology, Knowledge and Learning*. <https://doi.org/10.1007/s10758-025-09884-0>
- [2] Berber Sardinha, T. (2024). AI-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics*, 4(1), 100083. <https://doi.org/10.1016/j.acorp.2023.100083>
- [3] 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(3), 100068. <https://doi.org/10.1016/j.rmal.2023.100068>
- [4] Halliday, M. A. K. (1985). *An Introduction to Functional Grammar* (1 Edition). Edward Arnold. <https://doi.org/10.4324/9780203783771>
- [5] Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, 13(1), 18617. <https://doi.org/10.1038/s41598-023-45644-9>
- [6] Jiang, F. (Kevin), & Hyland, K. (2025a). Metadiscursive nouns in academic argument: ChatGPT vs student practices. *Journal of English for Academic Purposes*, 75, 101514. <https://doi.org/10.1016/j.jeap.2025.101514>
- [7] Jiang, F. (Kevin), & Hyland, K. (2025b). Rhetorical distinctions: Comparing metadiscourse in essays by ChatGPT

- and students. *English for Specific Purposes*, 79, 17-29.
<https://doi.org/10.1016/j.esp.2025.03.001>
- [8] Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior*, 114, 106553.
<https://doi.org/10.1016/j.chb.2020.106553>
- [9] Martin, J. R., & White, P. R. R. (2007). *The language of evaluation: Appraisal in English* (Paperback ed., 1. publ). Palgrave Macmillan.
- [10] Pütz, O., & Esposito, E. (2024). Performance without understanding: How ChatGPT relies on humans to repair conversational trouble. *Discourse & Communication*, 18(6), 859-868. <https://doi.org/10.1177/17504813241271492>
- [11] Qiu, W., & Pan, F. (2026). Comparing lexical complexity differences between ChatGPT-generated and human-written medical abstracts. *Scientometrics*, 131(8), 5049-5072.
<https://doi.org/10.1007/s11192-026-05701-2>
- [12] Turós, M., Soltész-Várhelyi, K., & Szűts, Z. (2026). Does the Thesis Still Make Sense? A Comparative Analysis of Scientific Essays Generated by Humans and Generative Artificial Intelligence. *Education Sciences*, 16(6), 920.
<https://doi.org/10.3390/educsci16060920>
- [13] Yao, G., & Liu, Z. (2025). Can AI simulate or emulate human stance? Using metadiscourse to compare GPT-generated and human-authored academic book reviews. *Journal of Pragmatics*, 247, 103-115.
<https://doi.org/10.1016/j.pragma.2025.07.018>
- [14] Zhang, M., & Crosthwaite, P. (2026). 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*, 64(2), 733-760. <https://doi.org/10.1515/iral-2024-0196>
- [15] Zhang, M., & Zhang, J. (2025). Disciplinary variation of metadiscourse: A comparison of human-written and ChatGPT-generated English research article abstracts. *Journal of English for Academic Purposes*, 76, 101540.
<https://doi.org/10.1016/j.jeap.2025.101540>