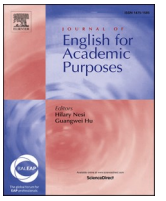
Journal of English for Academic Purposes 75 (2025) 101499

Contents lists available at ScienceDirect

Journal of English for Academic Purposes

journal homepage: www.elsevier.com/locate/jeap



Exploring the affordances of generative AI large language models for stance and engagement in academic writing

Zhishan Mo, Peter Crosthwaite *

School of Languages and Cultures, University of Queensland, Australia

ARTICLE INFO

Handling Editor: Dr Hilary Nesi

Keywords:
Stance
Engagement
English for academic purposes
Generative AI
Large language model
Academic writing

ABSTRACT

Large pre-trained models like ChatGPT demonstrate remarkable capabilities in generating coherent text across various domains, posing serious implications for teaching academic writing, given the potential for student plagiarism and reliance on software for developing writing skills. However, the linguistic properties and strategies these models employ remain largely unexplored. We investigate how three available large language models (LLMs) express stance and engage with readers in their writing, providing insights into their abilities to produce contextually appropriate and discipline-specific academic writing. 30 academic essays produced by each model were compared with those of human writers on identical topics using detailed prompts, before annotating each text for stance and engagement following Hyland's (2005) taxonomy. Results indicate that LLMs generally use a narrower and more repetitive range of stance and engagement features than human writers, with significant variation also across each LLM. Disciplinary use of stance and engagement is largely in line with human writing except for the philosophy discipline. Implications for teaching academic writing are discussed, particularly regarding identifying potential LLM-related plagiarism and inconsistencies in academic stance and engagement.

1. Introduction

The present study explores the production of *stance* and *engagement* (Hyland, 2005) in academic writing produced by human academic writers and that of three generative artificial intelligence (GenAI) models, specifically ChatGPT, MetaAI and ERNIEBot. The concepts of stance and engagement are a productive area of research in academic writing, primarily following the work of Hyland (2005) in determining how academic writers use specific lexical means to convey their stance on the claims they are making in the production of academic texts, as well as the means through which they engage with their academic readers in order to direct, question, appeal to shared knowledge, or specifically refer to the reader to suit a range of rhetorical purposes. Studies of stance and engagement in academic writing are numerous, spanning experimental, longitudinal, diachronic and disciplinary variables. These include diachronic corpora (Hyland & Jiang, 2016), corpus-based studies across disciplines (Hyland, 2005), the long-term effects of stance use after EAP instruction, and its influence on writing quality (e.g., Crosthwaite & Jiang, 2017).

The advent of generative artificial intelligence applications such as ChatGPT in late 2022 has fundamentally changed the way we consider both the process and product of academic writing within English for academic purposes (EAP) and English for research publication purposes (ERRP) contexts. Given the potential for generative AI applications to produce full texts supposedly constitutive of an academic register in response to users' natural language prompts, the implications of such technology for developing academic

* Corresponding author.
E-mail addresses: zhishan.mo@uq.net.au (Z. Mo), p.cros@uq.edu.au (P. Crosthwaite).

https://doi.org/10.1016/j.jeap.2025.101499
Received 12 November 2024; Received in revised form 26 January 2025; Accepted 1 March 2025
Available online 11 March 2025

writers, the academic writing process, assessment integrity and plagiarism, feedback and almost all other features of academic writing are only recently beginning to be explored in earnest. Studies comparing the academic register produced by GenAI models in comparison with that of human writers are now finally beginning to feature in the EAP literature (e.g. Jiang & Hyland, 2024; Oh & Lee, 2024). However, studies specifically targeting the use of stance and engagement features by GenAI applications, studies comparing the use of said features across human-produced and GenAI-produced academic writing, studies comparing said features by human/GenAI across disciplines, and even studies comparing the production of stance and engagement across different GenAI models are yet to come to the fore. By exploring the production of these features between humans and GenAI across disciplines and between GenAI models, the purpose of the present study is improving our understanding of the relative accuracy and shortcomings of current GenAI technologies in appropriating a human-like mastery of the academic register. Moreover, the purpose of comparing the stance and engagement in texts generated by the three LLMs is to evaluate their relative performance in replicating human-like academic discourse, identifying variations in their ability to employ key linguistic and rhetorical features. This comparison helps to uncover differences in how each model handles academic conventions, contributing to the understanding of model-specific strengths, weaknesses, and their suitability for academic writing tasks across disciplines.

The present study is significant in terms of being the first study (at the time of writing) to take a cross-disciplinary comparison of GenAI-produced stance and engagement following Hyland's (2005) framework, and also the only study (at the time of writing) comparing the production of stance and engagement features following Hyland's (2005) framework across three distinct LLMs. The implications of the findings of the study go beyond the exploration of simple lexical differences in that if we consider stance and engagement to be at the heart of 'successful' academic writing, any differences found across humans and AI, disciplinary areas or even between LLMs indicates AI is as yet unable to fully approximate human-like production of academic discourse. Implications for EAP provision, the development of LLMs, and those interested in the automated detection of AI-produced academic discourse are each investigated in the study's findings.

2. Literature review

EAP is the teaching and research of the academic variety of English needed for communication in academic contexts (Charles, 2012). Given the dominance of English as the academic *lingua franca* in an increasingly globalised world, EAP research and instruction has provided increasingly detailed explorations of the grammar, lexis, specificity and social dimensions of the academic register. EAP research has followed genre-based approaches including those in the Systemic Functional Linguistics (SFL) tradition (e.g. Halliday, 1994), or move-step analyses (e.g. Swales, 1990). From a social context, research has focused on the enculturation into academic disciplines provided by EAP instruction (e.g., Bazerman, 1988) or the gaps between so called 'native' and 'non-native' scholars in the linguistic hegemony of English as the dominant academic language (e.g. Tardy, 2004).

Research instruments exploring EAP language are primarily based around corpus analyses of collections of professional or learner academic texts, for example around the differences between spoken and written academic discourse (e.g. Biber, 2006), or on the specific lexical features of academic discourse and phraseology characterising disciplinary variation (e.g. Hyland, 2004, 2005). A particularly productive area of such research focuses on the reader/writer relationship as managed through the writers' use of *metadiscourse,* namely the selection of linguistic devices intended to convey the author's stance on the claims they make, and how they engage directly with the reader to suit specific rhetorical purposes.

2.1. Stance and engagement in academic writing

Hyland (2005) presented a model of linguistic features used to project the writers' stance in text, and devices used to presuppose the readers' active engagement with the written text as it unfolds. In Hyland's words:

> "put succinctly, every successful academic text displays the writer's awareness of both its readers and its consequences" (p.174).

Hyland's work arose from previous studies also exploring the evaluative nature of academic language including Halliday's (1994) 'attitude' dimension, Martin and White's (2005) 'appraisal' framework, and Biber and Finegan's earlier (1989) work on 'stance'.

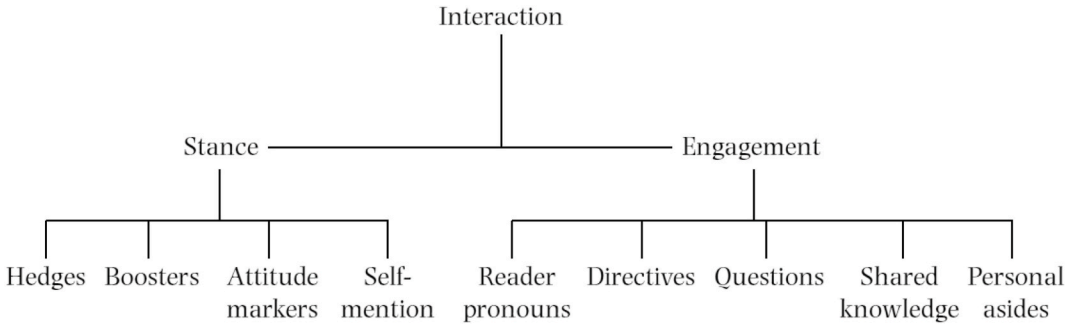


Fig. 1. Model of stance and engagement (taken from Hyland, 2005).

Taking a more specific view, Hyland (2005) proposed that evaluative meaning in text is composed between writers and readers with regards to the social dimension of the communities they comprise, "making rhetorical choices which evaluate both their propositions, and their audience" (p.175). Such evaluation is necessarily conveyed interactionally between the writer and reader. This concept is encapsulated in Fig. 1.

The writer's selection of linguistic devices to convey their attitudes towards their proposition is termed by Hyland (2005) as *stance*. Here, the writer can present themselves in their texts, emphasise their personal authority on any claims made, or move themselves (and their readers) with caution to any such claims. Stance features in academic discourse include the following:

Hedges: Devices including 'possible', 'might' and 'perhaps' are commonly used in academic writing to weak the author's commitment towards a claim to be made, avoiding risk of critique while acknowledging the current state of knowledge of the discipline.

Boosters: Expressions such as 'clearly' and 'obviously' are used to strengthen the author's commitment to the claims they are making, while marking the degree of shared knowledge held by the discipline in question.

Attitude markers: The author's use of expressions including 'unfortunately', 'hopefully' and 'surprisingly' signal an affective dimension to the claims the author chooses to produce in their text, demonstrating a presupposed sharing of values between the writer, reader and the discipline.

Self-mention: The use of pronouns including 'I', and 'we' when used to refer specifically to the writer is a key component of disciplinary variation in the authors' own intrusion into their texts.

Engagement then refers to how writers align the reader/writer relationship in text, recognising their readers explicitly, drawing their attention to certain statements or claims, and guiding them to interpret these claims in a particular manner. Engagement features in academic discourse include the following:

Directives: Terms including 'must', 'consider', 'imagine' and others directly instruct the reader to a specific action or conclusion by carrying out the writers' directive as they encounter it in text. These can be textual (e.g. 'see Lambert and Jones, 2005 … '), physical (e.g. 'set the slider to 300mm) or cognitive (e.g. 'consider Walker's claim that …) in nature.

Reader pronouns: The use of 'we', 'our' 'you', and 'us' mark a highly explicit method of drawing the reader directly into the author's discourse, directly signalling co-membership of the discipline in question between the writer and the reader.

Questions: Direct questions to the reader serve to increase writer/reader engagement as a form of dialogue, although in almost all cases are rhetorical (e.g. is it necessary to choose between ….)

Appeals to shared knowledge: Expressions such as 'as we all know … ' serve as a form of solidarity between the writer, the reader, and the current state of knowledge of the discipline, asking the reader to accept or recognise any claim as familiar territory.

Personal asides: Expressions such as 'by the way' mark the writers' intention to break up the flow of the text to address the reader directly. While not typical in the sciences, their use in the soft disciplines as a form of writer/reader dialogue is well-established (Hyland, 2005).

Studies exploring the use of stance and engagement in academic writing are numerous, spanning corpus-based investigations across disciplines (Hyland, 2005), diachronic corpora (Hyland & Jiang, 2016), collocation networks (Dong & Buckingham, 2018), and more. Beginning with Hyland's (2005) landmark study, Hyland investigated disciplinary variation in the use of stance and engagement features across 'soft' (e.g. arts and humanities, social science related) and 'hard' (e.g. physical, life science related) disciplines based on a corpus analysis of 240 research articles spanning three papers from each of ten leading journals across eight individual disciplines. The results show so-called soft disciplines make more frequent use of both stance and engagement features as writers in soft disciplines "take far more explicitly involved and personal positions" (p.187) than those in the hard disciplines. Related studies have also focused on linking specific stance and engagement discourse features to disciplinary practices (e.g., McGrath & Kuteeva, 2012). Other studies have sought to characterise the use of stance and engagement in other genres than the research article, e.g. 3-min thesis (3 MT) presentations (e.g., Qiu & Jiang, 2021), finding that unlike research articles, the hard disciplines make more use of interactional metadiscoursal language. Corpus studies have also investigated the changing use of stance over time in published research, with Hyland and Jiang's (2016) study exploring how academic writing is using more in the way of explicit stance features over the last few decades in an increasingly competitive and crowded world of academic publishing. Dong and Buckingham (2018) used corpora to map the collocation networks of stance expressions (e.g. the (results) *seem to* (affect)), noting several discipline-specific features as identified in these networks. Crosthwaite and Jiang (2017) explored the longitudinal development of stance in L2 EAP writing at the beginning, middle and end of EAP instruction at a university in Hong Kong SAR. Initial student writing is characterised by heavy use of boosters (e.g. '*obviously* the results *show* … ', 'it is *true* that'), self-mention and attitude markers (e.g. '*Unfortunately* ….'), yet becomes more cautious over time through a decrease in these forms and an increase in hedging use, in line with how the students received instruction on the value of making cautious, defendable claims on their EAP courses. Recent investigations (e.g. Jiang & Hyland, 2024) are also now turning to the investigation of stance and engagement in machine-generated academic discourse, to which we now turn the reader's attention.

2.2. Artificial intelligence and generative AI

Artificial intelligence (AI) is the umbrella term for a range of automated systems that carry out jobs that have historically been performed by people, using machine learning algorithms to learn from the system and get better over time (Luckin, 2017). AI encompasses a wide range of fields and technologies, permeating everyday life through devices like text prediction software and voice assistants. Artificial Neural Networks (ANNs), digital frameworks based on the neural structure of the human brain that learns from data to solve problems, are at the core of AI's imitation of human cognitive abilities (Yeralan & Lee, 2023). These applications evaluate language data and produce models that reflect language skill and usage by utilising natural language processing (NLP) in conjunction with rule-based systems or predictive/iterative AI models. These models are created through the application of machine learning and deep learning techniques. EAP uses artificial intelligence (AI) in several ways. These include automated essay scoring (AES) and automated writing evaluation (AWE) for grading written work, spoken dialogue systems and automated speech recognition (ASR) for assessing spoken language, machine translation (MT) for translating from other languages into English, and plagiarism detection.

Generative artificial intelligence (GenAI) refers to a specific set of artificial intelligence applications that employ sophisticated ANNs to generate novel, human-like outputs in response to a stimulus, like a query or instruction (Chan & Hu, 2023). Large Language Models (LLMs), which are the brains behind GenAI systems, may create original text, image, and audio material by training on enormous amounts of web-sourced data (Lim et al., 2023). Using sophisticated algorithms, LLMs generate contextually relevant and statistically likely phrases and sentences in addition to replicating previously learnt knowledge (Polverini & Gregorcic, 2024). This capability extends to complete phrases, where the model makes sure that newly added words blend in with the rest of the text to provide outputs that are grammatically correct, cohesive, and correctly punctuated (Yeralan & Lee, 2023).

2.3. Generative AI and academic writing

The rise of generative AI and large language models is a significant development in the EAP discipline, and has the potential to change the way EAP practitioners approach pedagogy, assessment, and curriculum design. GenAI systems, such as ChatGPT, can offer real-time feedback on grammar, vocabulary, and text coherence based on large amounts of data training. Additionally, these tools can assist users in identifying important concepts and recommend pertinent references (Kung et al., 2023). As evidence of GenAI's capabilities, ChatGPT successfully completed a first-class physics essay following the UK's higher education grading criteria in a recent study (Yeadon et al., 2023).

Notwithstanding this accomplishment, questions have been raised about the moral and ethical implications of academic plagiarism and AI authorship (Stokel-Walker, 2022). While several strengths of ChatGPT's function in education were highlighted in a recent SWOT analysis (Farrokhnia et al., 2024) including improved information access, decreased workloads, and personalized learning experiences, deficiencies such inadequate grasp of the subject matter, absence of critical evaluation, and inadequate use of higher-order cognitive abilities were also found. Other studies suggest GenAI poses concerns in the form of increased plagiarism, problems with academic integrity, and possible difficulties with equity and access (e.g. Stahl & Eke, 2024). For EAP, a continuing source of debate is the extent to which GenAI is used *as part of* the writing process rather than *being* the writing process, in that "different students and different academics "draw the line" in different places" (Rowland, 2023, p.T36). This 'line' can stretch from using AI for proofreading (e.g. spellcheckers or software e.g., Grammarly), to using AI to plan a structure, to using AI as a co-author, to students entirely adopting AI output as their own.

However, GenAI has several affordances for EAP teaching and instruction. Regarding feedback, for example, research has shown that GenAI can provide personalized feedback more relevant to students' needs than teachers' feedback. Chinese tertiary L2 English students found GPT3.5 and GPT4o′s written feedback more relevant than teachers' feedback, which focused mainly on accuracy (Li et al., 2024). The "AI + Teacher" model, which integrates teacher and GenAI feedback, has been successful in incorporating more co-produced feedback into revisions (Han & Li, 2024). GenAI applications can significantly enhance the editing and proofreading process if properly implemented and supervised by writing teachers; Japanese tertiary EAP students prefer GenAI-assisted editing and proofreading over student-led writing groups, for example, and EAP teachers appreciate the larger volume of feedback produced by ChatGPT (Allen & Mizumoto, 2024). However, studies have also found that written corrective feedback produced by ChatGPT varies considerably, even on the same prompt and same text across multiple chat sessions ((Lin & Crosthwaite, 2024).

Regarding GenAI's ability to accurately produce an academic register across disciplines, findings at the time of writing are mixed. Corpus studies have already identified several disparities between the academic register commonly produced by software such as ChatGPT and that of human academic writers, despite claims that generative AI can create a whole academic research article from scratch (Hsu, 2023). For instance, after using multidimensional analysis to evaluate register variation between academic papers generated by AI and human authors, Berber-Sardinha (2024) discovered "that, at present, AI's ability to capture the intricate patterns of natural language remains limited" (p.1). Additionally, Mizumoto et al. (2024) assembled a corpus of essays produced by ChatGPT in addition to a corresponding corpus of essays written separately by human L2 English writers in Japan. Using NLP and machine learning techniques to analyze the two corpora, significant differences were found across all linguistic features analysed (i.e., lexical diversity, clausal syntactic complexity, embedded syntactic complexity, complex nominals, modals, epistemic markers, and discourse markers). Following a Systemic Functional Linguistics approach, thematic choices and thematic progression patterns of L2 essays written by ChatGPT and humans were compared in Tang et al. (2024). Significant differences were noted across subtypes of textual, interpersonal, and marked topical themes, particularly regarding ChatGPT's use of concession signals, condition signals, modal adjuncts, and clause-initial conjunctive adjuncts, suggesting ChatGPT's production lacks human-like thematic development.

Focusing on lexical choices, Jiang and Hyland (2024) conducted a vocabulary-focused study wherein they compared the 3-word

lexical bundles used in a corpus of academic essays produced by ChatGPT with those of British students. Their findings revealed that ChatGPT employed a limited and repetitive range of these bundles, and there was less indication of bundles utilised to convey epistemic stance and authorial presence in comparison to human student writers. Similarly, Zhang and Crosthwaite (2025) examined the lexis and collocation patterns in academic essays written by ChatGPT-3.5 and a sample of graduate L2 academic essays. They discovered that while ChatGPT excels at producing texts with formal and complex vocabulary appropriate for academic and technical themes, L2 writers tend to concentrate on personal and social issues while using more varied and context-rich vocabulary. Accordingly, EAP educators are advised to focus on helping L2 writers communicate their unique experiences and obstacles through their academic writing rather than expecting more 'academic' but less human-focused production.

2.4. Summary

In summary, stance refers to how writers express their attitudes, judgments, and personal viewpoints, while engagement involves the ways writers connect with their readers and involve them in the argument or discussion. A significant body of research exists on stance and engagement, particularly in the context of academic writing. Similarly, there has been a growing interest in the role of GenAI in academic writing, exploring how these tools can assist or transform writing practices, including drafting, editing, and enhancing clarity. If EAP practitioners wish to meaningfully and ethically incorporate GenAI into their practice, understanding the limitations of LLM-produced academic discourse is vital in bridging the gap between what the technology can do and what novice academic writers still need to do to produce successful academic argumentation. Additionally, understanding these limitations will also help EAP practitioners to better detect unethical use of GenAI by students as part of the writing process. However, despite the wealth of research on these individual areas, there is relatively little exploration of the overlap between GenAI, stance, engagement, and academic writing. This gap presents an opportunity for further investigation into how GenAI might influence the way writers convey their stance and engage with readers, potentially reshaping traditional approaches to academic communication. Implications for EAP instruction and curricula, future LLM development, and automated detection of AI-produced academic discourse are numerous, and will be explored following the presentation of the results for each RQ, listed for the reader below.

RQ1: How does human- and GenAI-produced academic writing convey academic stance and engagement?
RQ2: Can GenAI-produced academic writing approximate human-like production of academic stance and register features across disciplines?
RQ3: To what extent do different GenAI models (i.e. ChatGPT, ERNIEBot, Meta AI) vary in their use of stance and engagement features in academic writing as compared with human academic writing and each other?

3. Data and methods

3.1. Data collection

As illustrated above, the present study compares academic essays generated by ChatGPT, ERNIE Bot and Meta AI with those of British students in tertiary education. For the latter, we used the British Academic Written English Corpus (BAWE, Alsop and Nesi, 2009), a collection of academic works written by UK students across different disciplines and study levels. 30 essays at levels 3 and 4 as identified in the corpus were extracted (corresponding to those written by final year undergraduate and masters' students respectively), with topics covering six disciplines (i.e. Classics, Archaeology, History, English, Philosophy and Linguistics).

For the three GenAI corpora, a series of contextual prompts was created to produce one essay for each topic following each requisite BAWE text's essay title, content and structure as closely as possible. The main method by which users engage with GenAI tools are prompts, which are basically queries or instructions that direct a language model to produce particular answers. Similar to honing search phrases for improved search engine results, users 'prompt engineer' these inputs to get more customised results. Template-based and regeneration-based quick formulation techniques were identified by Li et al. (2023). While template prompting makes use of pre-made templates that provide all necessary information, regeneration prompting reassesses and enhances original responses. The complexity of prompts varies; at the lower end of the spectrum are contextual prompts, role-based prompts, constraint-based prompts, and single-shot templates (simple instructions); at the higher end are regeneration-based nested and iterative prompts. Additionally, Li et al. (2023) identified three typical prompting template types: Example, Context, and Query. While Example and Query prompts are inappropriate for the task of academic essay writing, Context prompts can be enhanced with particular task information required to successfully generate an essay reflective of those found in BAWE. To do this, the researchers conducted a close reading of each BAWE text together, summarising the main points, critiques, observations and arguments, noting these in an excel file as notes from which to generate the context for each individual LLM prompt. Prior to final LLM data collection, a pilot test using these notes was conducted on each LLM platform for each individual essay, with the researchers together conducting an additional close reading to ensure this prompting approach generated texts as closely in line with the length, scope and register of academic writing as seen in the equivalent BAWE texts. Modifications were made to individual prompts where the researchers concluded there were notable differences between the GenAI-produced text and our notes from the close reading of the BAWE texts. In almost all cases identical prompts were used for each LLM, with only minor variations for certain texts. An example prompt for the topic of "E.M.Förster's preoccupation with

connection and his search fulfilment" can be seen below.

"I'm a Master student majored in English and my L1 is English, help me generate a 2000 words academic essay including citations, you need to address the title on "Discuss E.M.Forster's preoccupation with connection and his search fulfilment."
You can analyze E.M. Forster's focus on the theme of connection and his pursuit of fulfillment in his literary works. The essay should argue that Forster, a key figure in the Bloomsbury group, critiques the superficiality and repression of Victorian middle-class society, which he believes hinders true personal and social fulfillment. You can explore how Forster uses symbolism in his novels, such as Howards End and A Passage to India, to depict the struggles of characters as they seek deeper connections with others, nature, and their own roots. You can also discuss how Forster contrasts the alienating effects of modern urban life with the more meaningful connections possible in nature and through ancestry. You can further examine Forster's critique of societal conventions, intellectualism, and his reflections on his own life, particularly his homosexuality, which influence his portrayal of relationships and fulfillment. In the end, you can conclude that while Forster recognizes the difficulties in achieving lasting connections, he maintains a sense of hope for the human spirit's potential for fulfillment.
Please write the essay with the following structure and headings:
Introduction
Forster's Critique of Urban Life
Nature as a Source of Truth and Connection
Ancestry and Roots as Pathways to Fulfillment
Forster's Critique of Intellectualism and Romanticism
Forster's Personal Struggles and Their Influence on His Work
Spirituality and the Search for Cosmic Unity
Conclusion

The LLMs under investigation were ChatGPT 4.0, ERNIE Bot 4.0 and Meta AI. These LLMS were selected as the latest, state-of-the-art models available at the time of writing, offering distinct capabilities in text generation and analysis.

ChatGPT-4.0, developed by OpenAI and accessible through OpenAI's web interface https://chatgpt.com/, uses a transformer-based architecture with multimodal capabilities for rapid processing of both text and images. While the model's scale remains private, it was trained by using supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF), improving the model's performance to generate contextually appropriate and high-quality responses excelling in areas requiring complex reasoning, contextual understanding and high-quality content generation (OpenAI, 2023).

ERNIE Bot 4.0, China's leading large language model produced by Baidu, can be accessed by paying subscribers via https://yiyan.baidu.com/with a usage cap. The model is trained with over 260 billion parameters and incorporates a knowledge-enhanced architecture that combines both language understanding and knowledge integration. It is distinctive in the integration of both Chinese and English language capabilities, making it particularly effective for bilingual applications.

Meta AI, developed by Meta (Facebook) using the Llama 3.1 LLM at the time of access (released July 23rd, 2024) can be accessed through Meta's API and various partner platforms. Llama 3.1 uses a standard dense transformer model architecture, with the largest model featuring 405B parameters, capable of processing sequences up to 128K tokens in length within its context window. Compared with previous models, Llama 3.1 has demonstrated improved performance in complex reasoning tasks, creative writing, and code generation (Dubey et al., 2024).

Details of the four individual corpora as well as the 'GenAI' corpus of the three LLMs combined are shown in Table 1.

One consideration is the larger individual text lengths found in the BAWE data as compared with the three individual LLMs. Average text length for the BAWE essays was 2500 words, and despite repeated attempts at prompting for extended essays within the prompt used, the LLMs typically produced texts of between 800 and 1000 words in length. However, as the annotated stance and engagement features were extracted from both corpora to a normalised frequency of n per 1000 tokens, this mitigates discrepancies in corpus sizes.

3.2. Annotation of stance and engagement

Annotation of stance and engagement features was performed using UAMCorpusTool (version 3.3., O'Donnell, 2008 [1]). UAM is a freely available corpus annotation tool allowing for user-created corpus compilation using plain.txt files organised into a requisite folder structure for the human/LLM files. Each file was labelled in UAM for its 'author' (human/LLM), and its discipline. A taxonomy of stance and engagement features was generated in UAM using the same elements as found in Fig. 1 from Hyland (2005), allowing for ease of annotation (Fig. 2).

Annotation of each individual stance and engagement feature (e.g. hedge, booster, etc.) was carried out using the 'autocode' function in UAM. This involves the generation of a Corpus Query Language (CQL) regular expression whereby all instances of a given word are retrieved from the corpus and annotated according to its requisite stance/engagement feature (Fig. 3).

Lists of words for each stance and engagement feature were sourced from Hyland (2005). To expand the list further, the researcher also asked ChatGPT to generate additional entries for each category not found in Hyland (2005). ChatGPT was chosen for its ability to generate diverse linguistic examples through simple prompting. While potential biases may exist with GenAI output in this regard, all entries underwent rigorous manual review by the second author (an expert in research on stance and engagement) to determine their suitability for inclusion into the lists, in that each additional entry must have a metadiscursive function for stance and/or engagement within the context of academic writing in general, and in the specific passage(s) under review. This expert oversight served to mitigate potential biases in ChatGPT's output and ensure the appropriateness of all entries for the framework of stance and engagement in

[1] http://www.corpustool.com/download.html.

Table 1
Corpus sizes (tokens/words).

	Human	GenAI (combination of ChatGPT, Erniebot and MetaAI)	Chatgpt	Erniebot	MetaAI
Texts	30	90	30	30	30
Tokens	108200	157122	60123	43997	53002
Words	89662	130930	49014	38295	43621

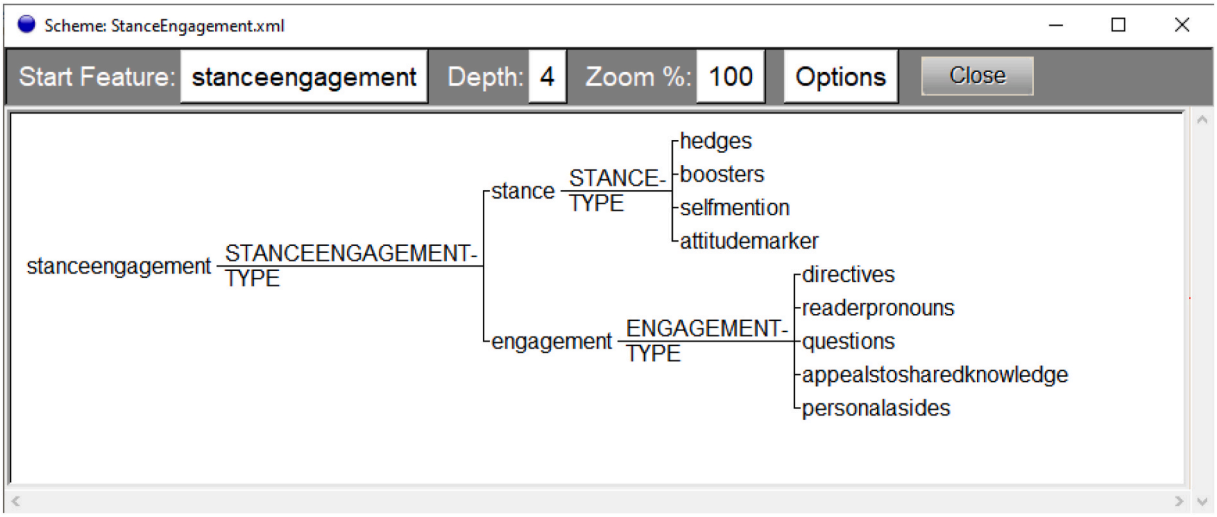


Fig. 2. Annotation framework for stance and engagement.

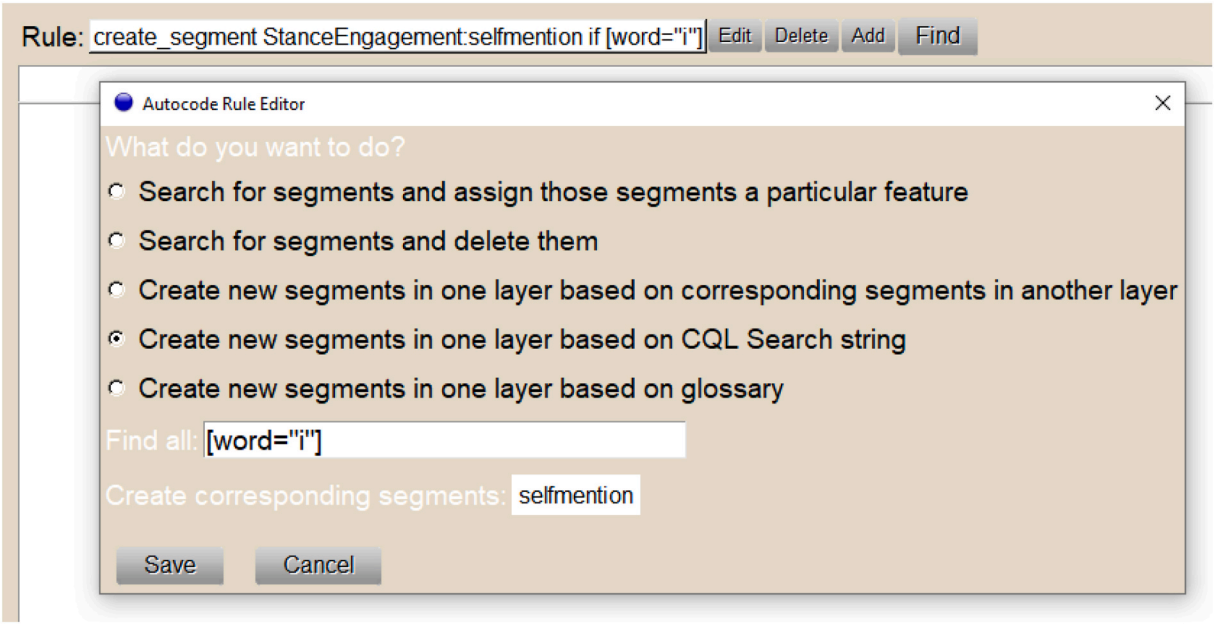


Fig. 3. Autocode function for 'self-mention' in UAMCorpusTool.

academic writing. The final list, including all of Hyland's original items and the new included GenAI-produced items is found in Appendix 1.

The next step involved determining if each annotated instance was in fact functioning as its annotated category, for example whether 'may' was being used as a hedge ("This *may* result in …") or not ("You *may* go outside"). This procedure involved a manual close reading of each text in UAM where each annotated feature is highlighted for the reader, making the process of identifying the

Table 2
Descriptives – AI vs Human production of stance and engagement devices.

			95% Confidence Interval Mean				
		Mean	Upper	Lower	Std. Deviation	Minimum	Maximum
stance	AI	13.745	15.202	12.288	6.955	1.24	32.9
	Human	18.885	22.09	15.68	8.583	2.74	40.36
engagement	AI	2.385	3.12	1.651	3.508	0	22.49
	Human	5.434	8.79	2.079	8.985	0	35.66

annotated features in each text much easier. All items not functioning under their assigned stance/engagement category had their annotations removed using a 'delete' function, with the two researchers conducting manual checking of all annotated items prior to deletion of those items.

Regarding coding reliability, as a first step, the first author conducted a first pass of each text, noting any annotations requiring possible removal. As a second step, the second author conducted a second pass, adding any additional required annotations while also noting annotations marked for deletion. Both authors then met to discuss each individual text, reaching consensus on the inclusion or deletion of annotated features over several sessions. In certain cases, individual words may function under two or more categories. For example, in Hyland (2005) the word 'clearly' functions both as a stance feature (booster) and an engagement feature (appeal to shared knowledge). UAM allows for individual words to receive multiple annotations. In their discussions, both authors discussed each multi-annotated item to determine if each individual annotation was valid given the context, removing any annotation if deemed inappropriate. Following this procedure, inter-rater agreement reached 100% over the entire dataset without the need for quantitative measures e.g., Cohen's Kappa, which is consistent with qualitative practices in similar research contexts (Miles et al., 2014).

3.3. Statistical analysis

UAMCorpusTool allows for the export of raw and normalised (n per 1000 tokens) frequencies of annotated features per individual corpus file as a.csv file. After transposing the data into long format, the JASP statistical package (version 0.19.00, https://jasp-stats.org/download/) was used to produce descriptive and inferential statistics for the production of stance and engagement features in human vs AI produced writing (RQ1), the discipline-specific production of stance and engagement features (RQ2), and comparison of the production of the three LLMs (RQ3). For inferential statistics, parametric analysis using ANOVA with post-hoc LSD tests, or non-parametric analysis using Kruskal-Wallis with post-hoc Mann-Whitney U tests, were conducted following Levene's test and Shapiro-Wilk assumption checks carried out in JASP, with the appropriate corrections made for significance in the event of multiple tests. Effect size measures used were Cohen's D for parametric tests and the rank-biserial correlation (ε^2) for non-parametric tests.

As an additional analytical procedure, the frequency of individual wordings for each stance/engagement category was retrieved via a corpus query in UAM. The frequency for each word, together with the total BAWE and GenAI corpus sizes were transferred into the Log-Likelihood (LL) Calculator (Rayson & Garside, 2000, http://ucrel.lancs.ac.uk/llwizard.html) to determine keyness across corpora. Typically, LL calculates statistical significance of keyness with values of 95th percentile; 5% level; $p < .05$; critical value = 3.84, 99th percentile; 1% level; $p < .01$; critical value = 6.63, 99.9th percentile; 0.1% level; $p < .001$; critical value = 10.83, 99.99th percentile; 0.01% level; $p < .0001$; critical value = 15.13. The present study mainly provides items with a significance value of $p < .0001$, which also offsets considerations for multiple tests. Additionally, a measure of effect size (%DIFF, Gabrielatos & Marchi, 2012) is provided for each analysis, indicating the proportion (%) of the difference between the normalised frequencies of a word in two corpora (or sub-corpora).

4. Results

This section presents the results of the analysis of stance and engagement features in light of the three research questions posed, in order.

4.1. RQ1 - Stance and engagement in human vs AI-produced academic writing

The following describes the relative use of stance and engagement devices used in the academic essays produced by human writers versus those of the average of the three AI tools (ChatGPT, ErnieBOT, and MetaAI) (RQ1). Separate analyses by discipline (RQ2) and by the AI tools themselves (RQ3) follow this description.

Table 2 and Fig. 4 describe the normalised frequencies per 1000 tokens of the use of stance and engagement devices between human and AI academic writing. The results suggest human writing features a higher frequency of both stance and engagement features as compared with AI writing. As the data did not meet parametric assumptions (Shapiro-Wilk $p < .005$), non-parametric Kruskal-Wallis inferential comparison reveals human academic writing is significantly more likely to contain both stance (H =

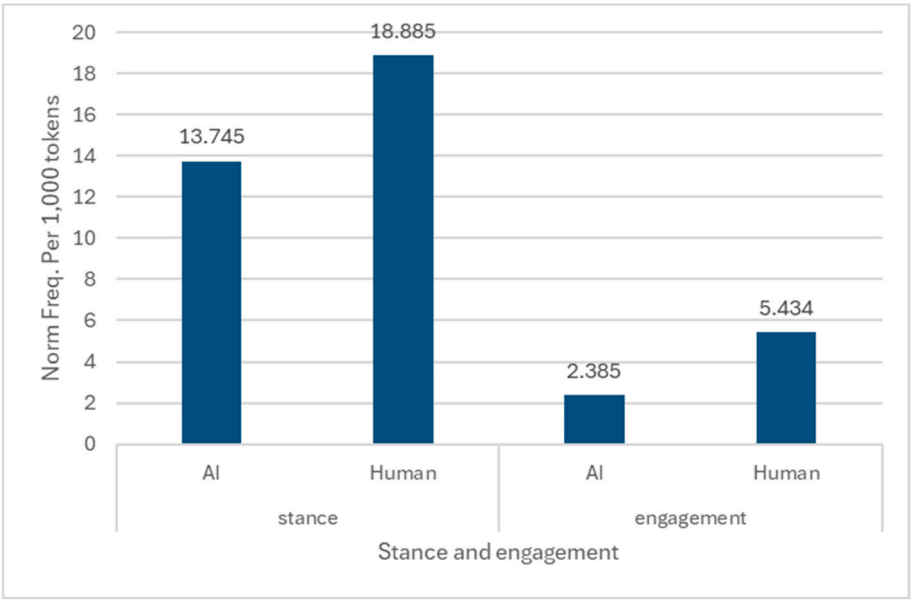


Fig. 4. Human vs. GenAI use of stance and engagement (n per 1000 tokens).

10.16, z = −3.18, p < .001, Rank ε^{2}[2] = 0.085) and engagement (H = 5.19, z = −2.28, p = .023, Rank ε^{2} = 0.040), with small effect sizes as shown in the Rank-Biserial correlation coefficient (Rank ε^{2}).

4.1.1. Stance

Table 3 and Fig. 5 describe the normalised frequencies (n per 1000 tokens) of the individual stance features produced in human vs AI-produced academic writing. Human writing appears to contain a higher frequency of hedges, boosters and self-mention, while AI-produced writing appears to contain a higher frequency of attitude markers. Individual Mann-Whitney U tests (adjusted p value = 0.0125 for 4 tests as per Bonferroni correction) reveal humans are significantly more likely to use hedges (U = 55.83, z = −2.58, p = .012, Rank ε^{2} = 0.053), and boosters (U = 51.91, z = −4.68, p=<0.001, Rank ε^{2} = 0.185), while results for self-mention and attitude markers were non-significant.

Turning now to the individual wordings, Table 4 describes the hedging devices more likely to feature in either human vs. AI written production following log-likelihood (LL) values of p < .0001 (LL > 15.18).

As Table 4 shows, human writers use a much wider range of frequently used hedging devices than those found in AI-produced writing. The use of 'would' represents the largest difference in hedging use between human and AI writing in terms of LL, with humans frequently using this term as a hedge to explore what others might claim.

(1) Where Marx *would* argue that ethnic groups are no more than 'national leftovers'

The largest effect size is found for the use of 'seems', frequently in the form 'seems to be'.

(2) Schopenhauer *seems* to be suggesting something he cannot be …

'Quite', 'probably', 'assumed', 'almost', 'apparently' and 'mainly' as hedges were not found at all in AI-produced writing as compared with human production. The only prominent hedging devices in AI-produced writing were 'often', with 287 hits compared with 60 in the human data, 'typically' which was rarely used by humans, and 'suggesting' which was never used in the human corpus as a hedge. The AI used 'often' with verbs to hedge frequent habitual claims.

(3) Early oriental scholars *often* portrayed these reasons as exotic …

Likewise, 'suggesting' was used within evaluative-that clauses to hedge claims made, e.g.

[2] Rank Biserial Correlation effect size benchmarks - small effect - 0.10 to 0.30, medium effect: 0.30 to 0.50, large effect: >0.50 (McGraw & Wong, 1992).

Table 3
Stance features used in Human vs AI academic writing.

Descriptive Statistics							
			95% Confidence Interval Mean				
		Mean	Upper	Lower	Std. Deviation	Minimum	Maximum
hedges	AI	7.03	8.103	5.957	5.124	0	25.78
	Human	9.341	11.281	7.4	5.196	0.55	23.57
boosters	AI	2.186	2.612	1.76	2.033	0	11.46
	Human	4.82	6.04	3.6	3.267	1.09	14.11
selfmention	AI	1.316	1.598	1.034	1.349	0	6.61
	Human	2.124	3.134	1.115	2.703	0	9.96
attitudemarker	AI	3.207	3.652	2.762	2.125	0	10.88
	Human	2.589	3.031	2.148	1.182	0.82	5.9

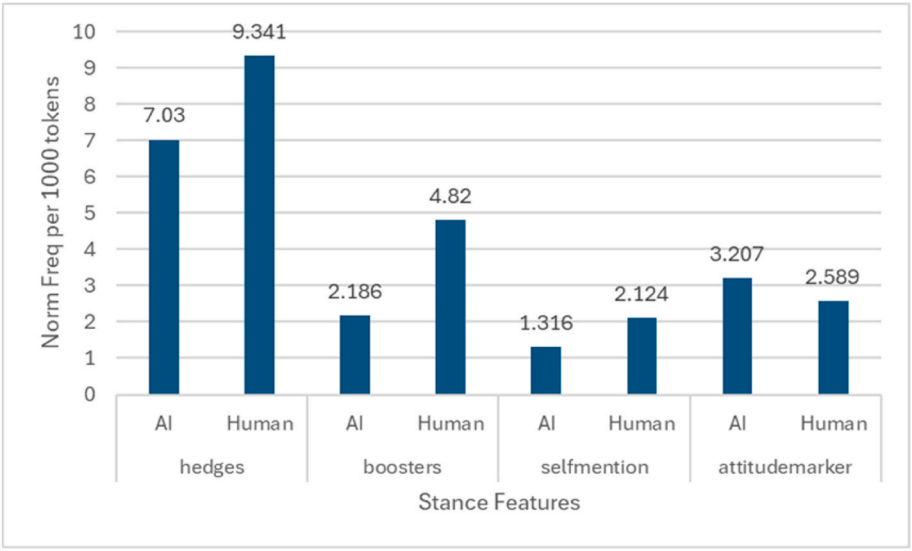


Fig. 5. Stance features used in Human vs AI-produced academic writing.

Table 4
Wordings of hedges (human vs GenAI).

Hedging Device	Frequency in human corpus	Frequency in AI corpus	Preferred in Human vs AI corpus	Log-likelihood	Effect size (%DIFF)[3]
would	146	37	Human	117.23	476.21
perhaps	47	2	Human	70	3331.62
seems	39	1	Human	61.91	5595.02
might	40	10	Human	32.41	484.1
quite	16	0	Human	28.81	NC[4]
seem	15	0	Human	27.01	NC
probably	14	0	Human	25.21	NC
assumed	12	0	Human	21.61	NC
appear	20	3	Human	21.33	873.51
suggested	31	10	Human	20.7	352.68
almost	11	0	Human	19.81	NC
apparently	9	0	Human	16.2	NC
mainly	9	0	Human	16.2	NC
often	60	287	AI	87.89	−69.47
typically	3	29	AI	15.75	−84.89
suggesting	0	21	AI	21.91	NC

[3] %DIFF indicates the proportion (%) of the difference between the normalised frequencies of a word in two corpora (or sub-corpora). A %DIFF value of 100 = 100% difference. There are no cut-off benchmark values because these vary depending on the corpora compared (Gabrielatos & Marchi, 2012).

[4] NC = Effect size not calculated as one of the corpora have no hits. %DIFF as an effect size is sensitive to no hit comparisons.

Table 5
Wordings of Boosters (human vs GenAI).

Boosting Device	Frequency in human corpus	Frequency in AI corpus	Preferred in Human vs AI corpus	Log-likelihood	Effect size (%DIFF)
in fact	28	0	Human	50.41	NC
clearly	33	2	Human	46.17	2309.43
indeed	32	3	Human	40.27	1457.61
never	22	2	Human	30.79	1506.29
believe	19	3	Human	19.81	824.83
actually	13	1	Human	17.25	1798.34
thought	9	0	Human	16.20	NC
shows	28	11	Human	15.49	271.70

Table 6
Wordings of self-mention (human vs GenAI).

Self-mention Device	Frequency in human corpus	Frequency in AI corpus	Preferred in Human vs AI corpus	Log-likelihood	Effect size (%DIFF)
I	76	25	Human	49.88	343.92
this paper	11	0	Human	19.81	NC
my	14	1	Human	18.90	1944.37
this essay	1	38	AI	32.14	−96.16

(4) Both films use their fantastical elements to critique the real-world dangers associated with nuclear technology, *suggesting* that the unchecked spread of these weapons could lead to unforeseen and disastrous consequences.

Table 5 describes the wording comparisons for individual boosters.

As with hedges, human-produced academic writing contains a wider range of frequently used boosting devices, with AI-produced writing using a lower frequency of all boosting devices with no single booster more likely to be found in that corpus. The highest effect size is found for the use of 'clearly', which human writers used frequently as a booster together with verbs e.g. illustrated.

(5) These changing ethics were *clearly* illustrated in the literature of the time …

Table 6 describes the wording comparisons for self-mention.

Human writers made more frequent use ($p < .0001$) for the use of 'I', 'this paper' and 'my' as self-mention. The largest effect size is found in the use of 'my', which humans use when describing personal stories or situations.

(6) I will be looking at what aspects of pronunciation should be taught in *my* teaching situation

However, AI-produced writing frequently features the use of 'this essay' as a self-mention, almost always in the phrase 'this essay will explore … ' or 'this essay examines … '.

(7) *This essay* examines the phenomenon of 'bomb cinema'

Table 7 describes the wording comparisons for attitude markers ($p < .001$, $LL > 10.80$)

It is apparent that AI-produced writing frequently signals attitude using a single device - 'significant'.

(8) Christian historiography played a *significant* role in shaping the preservation of ….

Human-produced writing, on the other hand, uses 'clearly' to signal attitude towards claims made.

(9) This initial sequence is *clearly* subjective and the camera itself …

The use of 'interesting', 'obvious', 'interestingly' and 'unfortunately' as attitude markers also punctuate human-produced academic writing but did not feature in AI-produced academic writing at all.

4.1.2. Engagement

The data in Table 8 and Fig. 6 suggest that the use of direct engagement in both human and AI-produced academic writing is of a low frequency, which is in line with academic writing norms (Hyland, 2005). The exception to this is found in the use of reader pronouns, although subsequent Mann-Whitney U-tests (adjusted p value = 0.0125 for 4 tests) found no significant difference between human vs AI produced writing in the use of such pronouns ($p = .154$). Significant differences were found however in the use of appeals to shared knowledge ($U = 52.15$, $z = -7.33$, $p < .001$, Rank $\varepsilon^2 = 0.452$) with a strong effect size. These barely featured in AI-produced

Table 7
Wordings of attitude markers (human vs GenAI).

Attitude marking Device	Frequency in human corpus	Frequency in AI corpus	Preferred in Human vs AI corpus	Log-likelihood	Effect size (%DIFF)
significant	26	213	AI	104.62	−82.18
clearly	29	2	Human	39.47	2017.38
interesting	11	0	Human	19.81	NC
obvious	6	0	Human	10.80	NC
interestingly	6	0	Human	10.80	NC
unfortunately	6	0	Human	10.80	NC

Table 8
Engagement features used in Human vs AI academic writing.

Descriptive Statistics							
			95% Confidence Interval Mean				
		Mean	Upper	Lower	Std. Deviation	Minimum	Maximum
directives	AI	0.355	0.485	0.225	0.621	0	2.91
	Human	0.413	0.689	0.137	0.739	0	3.67
readerpronouns	AI	2.025	2.688	1.362	3.164	0	21.16
	Human	4.617	7.755	1.479	8.404	0	33.7
questions	AI	0	NaN	NaN	0	0	0
	Human	0.009	0.028	−0.01	0.051	0	0.28
appealstosharedknowledge	AI	0.006	0.018	−0.006	0.058	0	0.55
	Human	0.396	0.556	0.236	0.427	0	1.44

[a] All values are identical.

writing but did feature in human-produced writing.

No significant differences of wordings were found for directives or questions. Regarding reader pronouns, Table 9 describes differences found at the $p < .0001$ level.

The use of 'we' as a reader pronoun is significantly more likely to be found in human-produced academic writing, often used in conjunction with directives.

(10) … and it is to this that *we* now turn *our* attention.

Similarly, the use of 'us' as reader pronouns is also a significant feature of human-produced writing, addressing solidarity with the reader.

(11) Class is more fluid than this attack would have *us* believe.

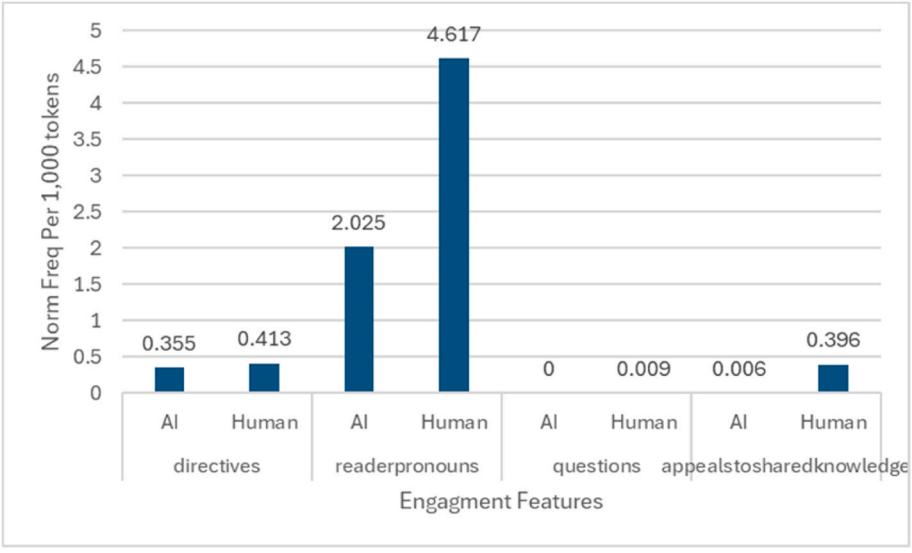


Fig. 6. Engagement features (human vs GenAI, n per 1000 tokens).

Table 9
Wordings of self-mention (human vs GenAI).

Reader pronoun Device	Frequency in human corpus	Frequency in AI corpus	Preferred in Human vs AI corpus	Log-likelihood	Effect size (%DIFF)
- we	205	148	human	43.40	102.27
- us	62	23	human	36.38	293.64
- one	55	16	human	39.95	401.97

Table 10
Normalised frequencies of stance features in AI vs Human writing by discipline.

Descriptives - stance						
Human vs AI	Discipline	N	Mean N per 1000 words	SD	SE	Coefficient of variation
AI	Archaeology	15	15.29	6.853	1.769	0.448
	Classics	15	10.34	4.384	1.132	0.424
	English	15	10.33	4.671	1.206	0.452
	History	15	12.31	7.428	1.918	0.603
	Linguistics	15	11.65	3.987	1.030	0.342
	Philosophy	15	22.51	5.629	1.453	0.250
Human	Archaeology	5	22.60	5.442	2.434	0.241
	Classics	5	9.240	4.159	1.860	0.450
	English	5	16.62	5.317	2.378	0.320
	History	5	13.31	1.115	0.499	0.084
	Linguistics	5	21.56	7.688	3.438	0.357
	Philosophy	5	29.95	7.592	3.395	0.253

Finally, the use of 'one' as a reader pronoun is a significant feature of human-produced writing, often used to exemplify actions for an imagined reader to follow.

(12) If *one* understands and executes the method of reduction properly, *one* can be sure that *one* sees the essences of ….

4.2. RQ2 – Stance and engagement features in human vs AI writing by discipline

Table 10 describes the normalised frequencies of stance features in Human vs AI writing by discipline.

As parametric comparison of human vs AI-produced stance was found to meet assumption checks (Levene's test $p = .163$), ANOVA analysis of the interaction of human vs AI-produced stance features and discipline was not found to be significant ($p = .068$). Post-hoc tests (LSD, with corrections for 66 tests) also found significant differences between the use of stance features in human vs. AI produced writing for each comparable discipline (i.e. human classics vs. AI classics, etc.). This suggests that, overall, the AI-produced texts are using a comparable frequency of stance features in line with human-produced disciplinary norms.

Regarding engagement, Table 11 compares the mean normalised frequencies of engagement features produced in human and AI writing by discipline.

As parametric comparison of human vs AI-produced stance was found to meet assumption checks (Levene's test $p = .87$), ANOVA analysis of the interaction of human vs AI-produced stance and discipline was found to be significant ($p < .001$). Post-hoc tests (LSD, with corrections for 66 tests) found differences between the use of engagement features in human vs. AI produced writing for the Philosophy discipline (mean difference = −15.85 for AI as compared with human data, $p < .001$, Cohens' D = -0.31[3]). Exploring further, this difference is found in the use of reader pronouns (MD = −15.30, $p < .001$, Cohen's D = −0.38) and appeals to shared knowledge (MD = −0.526, $p < .001$, Cohen's D = −0.81). AI-produced essays for the philosophy discipline are significantly less likely to feature reader pronouns or appeals to shared knowledge, which are a significant and frequent feature of human-produced essays in this discipline (see also Hyland, 2005), as shown in the following example.

(13) *Clearly,* if *we* consider the idea of a pen, not every member of the community will have the same image of a pen. What *we* are looking for here is …

4.3. RQ3 – Differences in the use of stance and engagement between the three AI models

Table 12 describes the use of stance and engagement features across the three AI models. The human data is also shown for comparison.

As parametric assumptions were met (Levene's test = p = .107), ANOVA analysis of the use of stance features across the four 'authors' was significant (F = 4.78, df = 3, p.004, n2 = 0.110), with post-hoc test revealing significant differences between human-

[3] Effect size benchmarks = small (d = 0.2), medium (d = 0.5), and large (d = 0.8) based on benchmarks suggested by Cohen (1988).

Table 11
Engagement features by discipline.

Human vs AI	discipline	N	Mean	SD	SE	Coefficient of variation
AI	Archaeology	15	2.45	2.987	0.771	1.219
	Classics	15	0.741	0.882	0.228	1.191
	English	15	1.843	2.595	0.67	1.408
	History	15	0.99	1.465	0.378	1.48
	Linguistics	15	2.099	1.63	0.421	0.777
	Philosophy	15	6.19	6.009	1.552	0.971
Human	Archaeology	5	1.344	0.667	0.298	0.496
	Classics	5	0.324	0.48	0.215	1.482
	English	5	3.494	1.402	0.627	0.401
	History	5	2.792	1.563	0.699	0.56
	Linguistics	5	2.61	1.59	0.711	0.609
	Philosophy	5	22.042	12.49	5.586	0.567

Table 12
Stance and engagement features across the LLMs.

				95% Confidence Interval Mean		
		Valid	Mean	Upper	Lower	Std. Deviation
stance	Ernie	30	15.363	18.275	12.451	7.799
	GPT	30	13.957	16.147	11.768	5.864
	Human	30	18.885	22.09	15.68	8.583
	Meta	30	11.915	14.476	9.355	6.857
engagement	Ernie	30	2.686	4.444	0.928	4.708
	GPT	30	1.258	2.032	0.483	2.074
	Human	30	5.434	8.79	2.079	8.985
	Meta	30	3.212	4.345	2.08	3.033

produced writing and that of MetaAI, with a large effect size (MD = 6.90 in favour of human writing, p = .002 corrected for 6 tests, Cohens' D = 0.949). This suggests the use of stance features in the writing produced by MetaAI was far less than that of human writers, while the use of these features by GPT and ErnieBOT was comparable. Exploring further, significant variation (F = 2.87, df = 3, p = .039, n2 = 0.069) was found in the use of hedges between human and MetaAI (MD = 3.78, p = .029, Cohens' D = 0.743). Likewise, significant variation (F = 9.51, df = 3, p < .001, n2 = 0.197) was found in the use of boosters between human and MetaAI (MD = 2.86, p < .001, Cohens' D = 1.191), ErnieBOT and human (MD = −2.2, p = .003, Cohens' D = −0.916) and GPT and human (MD = −2.84, p < .001, Cohens' D = −1.183). This suggests all three AI models failed to produce boosters to the same extent as human writers.

Turning now to engagement, as parametric assumptions were not met, Kruskal-Wallis tests revealed significant variation across the four 'authors' in terms of the use of engagement (H = 15.76, p = .001). Post-hoc tests (Dunn) revealed differences between GPT and human use of engagement (U = 42.13, z = −3.43, p = .003) and GPT and MetaAI (U = 42.13, z = −3.29, p < .001). Exploring further, this variation appears to be sourced in the use of reader pronouns (H = 14.39, p = .002) and appeals to shared knowledge (H = 53.92, p < .001), with variation in the use of reader pronouns between GPT and human (U = 43.90, z = −2.74, p = .003) and GPT and MetaAI (U = 43.90, z = −3.45, p < .001), and variation in the use of appeals to shared knowledge between ERNIEBot and human (U = 51.50, z = −6.10, p < .001), GPT and human (U = 53.45, z = −5.75, p < .001) and human and MetaAI (U = 85.50, z = −6.13, p < .001), with humans tending to use a higher frequency of both engagement types than the AI models, and with some variation between the AI models as well.

5. Discussion

Overall, the study has found significant variation in the production of stance and engagement features between human- and machine-produced academic discourse, including disciplinary variation in the use of said features, as well as variation across the three large language models (LLMs) used in the present study. We now discuss the findings for each RQ in turn.

5.1. RQ1: How does human- and GenAI-produced academic writing convey academic stance and engagement?

The present study revealed human academic writers a) use a significantly higher frequency of lexical stance and engagement features compared with the writing produced by GenAI, and b) that human writers use a wider range of stance and engagement features as compared with GenAI-produced academic discourse. GenAI-produced discourse, by contrast, tends to feature fewer stance and engagement features, repeating a narrower set of said features more often. This is evidenced in the increased use of both hedges (including terms e.g. 'would', 'perhaps' and 'seems') and boosters (including terms e.g. 'indeed, 'in fact' and 'clearly') in human-produced academic writing, while LLM-produced writing appeared to contain a significantly higher frequency of attitude markers – however, that appears restricted to heavy use of the term 'significant' within LLM academic production. The findings reflect

fundamental differences in how humans and GenAI construct academic arguments and engage with readers. Human writers likely draw on a richer repertoire of linguistic strategies, developed through experience and training, to align with the conventions of academic discourse. This allows them to employ a diverse range of stance features—like hedges and boosters—that demonstrate nuanced, context-sensitive argumentation. In contrast, the narrower use of these features in GenAI writing may result from the models' training data and algorithms. LLMs generate text based on statistical patterns in their training data rather than a contextual understanding of rhetorical or disciplinary conventions. This could lead to over-reliance on frequently occurring terms like "significant" for attitude marking, without the same variety or contextual adaptation evident in human writing. The repetition of a narrower set of features suggests limitations in the LLMs' ability to mimic the subtle, dynamic choices human writers make to balance authority, caution, and engagement in their texts.

These findings for LLMs are at odds with diachronic corpus studies of stance features that reveal human writers have increased their use of such features in academic writing over the past decades (Hyland and Jiang, 2016). However, the findings align almost exactly with Jiang and Hyland's (2024) investigation of 3-word lexical bundles in ChatGPT-produced writing, in that both that study and the present found LLM-produced academic writing contained a lower frequency and narrower set of expressions used to convey epistemic stance. While inferential statistics did not support a finding that humans used more self-mention to intrude into their texts than LLMs in the present study as was found in Jiang and Hyland (2024), the present study did however find human writers produced a higher frequency and larger range of reader pronouns as an engagement device as compared with LLM-produced writing, which is in line with previous findings. We concur with Berber-Sardina (2024) "that, at present, AI's ability to capture the intricate patterns of natural language remains limited" (p.1) when comparing academic papers generated by AI and human authors, while our findings also support similar human/LLM comparisons of academic discourse e.g., Mizumoto et al. (2024), and Yang et al. (2024).

5.2. RQ2: Can GenAI-produced academic writing approximate human-like production of academic stance and register features across disciplines?

In contrast to the finding from RQ1, the resulting analyses for RQ2 paint a positive picture for LLM-produced stance and engagement across disciplines, in that LLM production of these features was not found to be significantly different from that of human production across 5 of the 6 disciplines under investigation. This suggests that the training data underpinning current LLM technology has sufficient coverage of the humanities-focused disciplinary areas used in the present study.

The only difference between human and LLM production is found for the philosophy discipline, which is characterised in human production by very frequent use of reader pronouns and appeals to shared knowledge. Our findings concur exactly with Hyland's (2005) corpus-based investigation of stance and engagement features in philosophy as compared with other soft (and even hard) disciplines, where philosophy accounted for the highest frequency of both reader pronouns and appeals to shared knowledge, as also found in the present study. As "rhetorical practices are inextricably related to the purposes of the disciplines" (Hyland, 2005, p. 187), our findings both confirm Hyland's (2005) findings for human writers while at the same time problematise the academic production of LLMs regarding true disciplinary specificity. Accounting for the mismatch between human and LLM production for philosophy could indicate a lack of relevant training data for this discipline for each LLM or could be an artifact of the finding that LLMs produce a lower frequency of stance and engagement markers in general, although we do not have firm evidence in either direction.

5.3. RQ3: To what extent do different GenAI models (i.e. ChatGPT, ERNIEBot, MetaAI) vary in their use of stance and engagement features in academic writing as compared with human academic writing and each other?

The results for the three LLMs (as compared with both human production and each other) reveal significant variation across each model and humans for both stance and engagement. Regarding stance, MetaAI-produced academic discourse least approximated human production as compared with ChatGPT and ERNIEBot in terms of overall frequency, characterised mainly by a lack of hedging devices in MetaAI's production. All three models however struggled to approximate human-like production of boosters, suggesting that each LLM erred toward caution in their academic production and withheld opportunities to stamp their authority on any claims made. Regarding engagement, each LLM struggled to reproduce a human-like frequency of reader pronouns and appeals to shared knowledge. While this is likely an artifact of the previous finding for RQ2 regarding the impact of the philosophy discipline on the production of these forms, it is still noteworthy to report the mismatch between the LLMs and human writers in this respect.

The differences in stance and engagement across the three LLMs likely stem from variations in training data, model design, and the inherent limitations of AI in replicating human rhetorical strategies. As the data suggest, all three models struggle with boosters and engagement features due to their reliance on statistical patterns over rhetorical intent. AI's risk-averse tendencies further reduce strong claims, and engagement features, reflecting human-centric interaction, remain difficult for LLMs to replicate without situational awareness. These factors highlight the models' limitations in producing truly human-like academic discourse.

6. Conclusion

The present study has investigated the use of stance and engagement features in the academic writing produced by human writers and those of three LLMs. This study contributes theoretically by advancing our understanding of stance and engagement in academic writing, highlighting how these features are shaped by the relationship between linguistic, rhetorical, and disciplinary conventions across human and AI-produced texts. By revealing the limitations and patterns of generative AI in replicating these features, the study provides a framework for evaluating stance and engagement in LLM-produced academic discourse and emphasises the need to refine

theoretical models of stance and engagement to account for emerging AI-driven text production. Taken together, this study's findings support the following quote from Jiang and Hyland's (2024) study, namely:

> "that ChatGPT is less adept at injecting the text with a sense of personal perspectives and persuasive interactions that are typically valued in argumentative writing. This takes nothing away from our positive assessment of the essays it generated nor are we undervaluing the obvious power and affordances of ChatGPT for writing assistance. It just does not do this in the same way, or as effectively, as human writers" (p.14).

It is apparent that while GenAI is a game-changing tool to produce academic discourse, the implications of the present study for EAP and future LLM production are numerous. For students, those wishing to use GenAI during the writing process should be aware of the differences in how LLMs and more advanced human academic writers produce stance and engagement. The relative lack of stance and engagement production in LLM-written text risks the author - if using GenAI to create academic discourse - losing an essential component of academic argumentation, namely the ability for authors to either weaken potential claims to mitigate criticism, stamp their authority on any claims made, insert themselves into the texts, convey their attitude towards the subject matter through lexical means, or engage with the reader through a variety of accepted academic devices for doing so. This may leave student writing less personal, more 'robotic', and, in a sense, less truly 'academic', at least in terms of the arts and humanities-related disciplines investigated in the present study.

Regarding EAP practitioners and implications for pedagogy, the findings of the present study should more greatly emphasise the importance of explicit training for students in the lexical devices used by academic writers for stance and engagement. These already feature in several published initiatives (e.g. Hyland, 2005) but are needed perhaps more than ever in the GenAI era. The study's findings continue to emphasise the value of corpus linguistic investigations of academic discourse in general and those comparing human vs. GenAI produced discourse in particular, in that the evidence base corpus linguistics provides can serve to dissuade EAP sceptics that instruction in EAP is no longer required in light of GenAI technology. Incorporating corpus findings into EAP instruction are still valuable in providing evidence-based examples of how stance and engagement features are used in academic writing, allowing students to better understand and replicate actual disciplinary conventions rather than LLM-produced ones. There is obviously some way to go before LLM technology can exactly approximate human-produced academic lexis for stance and engagement, although, in fairness, it is still doing an excellent job. Considering this, EAP practitioners should guide students on how to use generative AI as a supplementary tool while prioritising the development of their independent academic writing skills to avoid over-reliance on AI-generated content.

Regarding LLM development, it is apparent that not all LLMs are created equal, and students' selection of specific LLMs according to their needs or preferences may impact the quality of the academic output they eventually incorporate into their writing. MetaAI stands out among the three LLMs as 'less academic' at least in terms of its production of stance and engagement features. While we are still yet to fully understand difference across LLMs beyond calculation of training data sizes and sources or number of machine learning parameters (i.e., is bigger always better?), the findings of the present study suggest that such training data must do better at capturing disciplinary variation in the use of stance and engagement. Future research should also investigate whether competitor LLMs (e.g. Claude, Gemini) fare any better than the three LLMs surveyed in the present study.

Regarding the study's limitations and opportunities for future research, it may be acknowledged that the number of texts for analysis ($n = 30$) could have been increased, with only five texts analysed per discipline. The manual nature of checking each individual annotation to determine whether it was being used according to its stance/engagement function is a time-consuming process (some human texts were over 3000 words in length) involving full inter-coder agreement, and so the decision was taken to limit the number of texts to 30 for both humans and individual LLMs. Future studies could use AI to automatically annotate stance and engagement features over a much larger body of texts. However, our initial pilot attempts at doing so proved problematic, with significant individual variation within and across individual LLMs on the same texts using the same prompts, necessitating a smaller, manual approach. The focus on arts and humanities-related (so-called 'soft') disciplines only is also a limitation of the present study. Future research needs to investigate so-called 'hard' disciplines (i.e. physical, life sciences) to determine the relative performance of humans/LLMs on these disciplines given reported variation across soft/hard disciplines in prior research (e.g. Hyland, 2005), while also providing a foundation for refining AI models to better serve academic purposes across different disciplines.

Another limitation that applies to many GenAI-related published studies is that the nature of the prompts used in the study will almost certainly impact the GenAI output. The greatest care was taken to ensure compatibility between the human/GenAI corpora through pilot prompting; however different results may well have been obtained if a request for "please use a range and high frequency of lexical devices for stance and engagement in your response" was provided in the prompt itself. While, in our opinion, this request would be unlikely to feature in most typical novice EAP students' prompts for academic writing, more advanced students who had taken EAP training may well choose to incorporate such a parameter into their prompt, with different results. Of course, the inherent lack of replicability of GenAI-produced output across the same prompts, same texts but carried out in different prompting/chat sessions on different days may also affect the replicability of the present study's findings, but this criticism can also be extended to much research investigating GenAI production at present.

Despite the limitations of the present study, however, we are confident that the present study adds significantly to the body of research on GenAI-produced academic writing, with clear pedagogical implications for EAP students, practitioners and LLM developers.

CRediT authorship contribution statement

Zhishan Mo: Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Peter Crosthwaite:** Writing – original draft, Visualization, Supervision, Formal analysis, Conceptualization.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jeap.2025.101499.

References

Allen, T. J., & Mizumoto, A. (2024). ChatGPT over My Friends: Japanese English-as-a-Foreign-Language learners' preferences for editing and proofreading strategies. *RELC Journal*. https://doi.org/10.1177/00336882241262533

Alsop, S., & Nesi, H. (2009). Issues in the development of the British academic written English (BAWE) corpus. *Corpora, 4*(1), 71–83. https://doi.org/10.3366/e1749503209000227

Bazerman, C. (1988). *Shaping written knowledge*. University of Wisconsin Press.

Berber-Sardinha, T. (2024). AI-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics, 4*(1), Article 100083. https://doi.org/10.1016/j.acorp.2023.100083

Biber, D. (2006). *University language*. John Benjamins.

Biber, D., & Finegan, E. (1989). Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text - Interdisciplinary Journal for the Study of Discourse, 9*(1), 93–124.

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education, 20*(1), 1–18. https://doi.org/10.1186/s41239-023-00411-8

Charles, M. (2012). English for academic purposes. *The Handbook of English for Specific Purposes*, 137–153. https://doi.org/10.1002/9781118339855.ch7

Cohen, J. (1988). *Statistical power analysis for the Behavioral sciences*. New York: Routledge.

Crosthwaite, P., & Jiang, K. (2017). Does EAP affect written L2 academic stance? A longitudinal learner corpus study. *System, 69*, 92–107.

Dong, J., & Buckingham, L. (2018). The collocation networks of stance phrases. *Journal of English for Academic Purposes, 36*, 119–131.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., & Rodriguez, A. (2024). The Llama 3 Herd of models. ArXiv.org https://arxiv.org/abs/2407.21783.

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2024). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education & Teaching International, 61*(3), 460–474. https://doi.org/10.1080/14703297.2023.2195846

Gabrielatos, C., & Marchi, A. (2012). Keywords: Appropriate metrics and practical issues. In *CADS International Conference, Bologna, Italy, 13-15 September 2012*. Retrieved from https://repository.edgehill.ac.uk/4196.

Halliday, M. A. K. (1994). *An introduction to functional grammar* (2nd ed.). Edward Arnold.

Han, J., & Li, M. (2024). Exploring ChatGPT-supported teacher feedback in the EFL context. *System, 126*, Article 103502. https://doi.org/10.1016/j.system.2024.103502

Hsu, H.-P. (2023). Can generative artificial intelligence write an academic journal article? Opportunities, challenges, and implications. *Irish Journal of Technology Enhanced Learning, 7*(2), 158–171. https://doi.org/10.22554/ijtel.v7i2.152

Hyland, K. (2004). Disciplinary interactions: Metadiscourse in L2 postgraduate writing. *Journal of second language writing, 13*(2), 133–151.

Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse. *Discourse Studies, 7*(2), 173–192. https://doi.org/10.1177/1461445605050365

Hyland, K., & Jiang, F. (2016). Change of attitude? A diachronic study of stance. *Written Communication, 33*(3), 251–274. https://doi.org/10.1177/0741088316650399

Jiang, F., & Hyland, K. (2024). Does ChatGPT argue like students? Bundles in argumentative essays. *Applied Linguistics*. https://doi.org/10.1093/applin/amae052. *amae052*.

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health, 2*(2), Article e0000198. https://doi.org/10.1371/journal.pdig.0000198

Li, J., Huang, J., Wu, W., & Whipple, P. B. (2024). Evaluating the role of ChatGPT in enhancing EFL writing assessments in classroom settings: A preliminary investigation. *Humanities and Social Sciences Communications, 11*(1). https://doi.org/10.1057/s41599-024-03755-2

Li, Y., Sha, L., Yan, L., Lin, J., Raković, M., Galbraith, K., Lyons, K., Gašević, D., & Chen, G. (2023). Can large language models write reflectively. *Computers and Education: Artificial Intelligence, 4*, Article 100140. https://doi.org/10.1016/j.caeai.2023.100140

Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A Paradoxical perspective from management educators. *International Journal of Management in Education, 21*(2), Article 100790. https://doi.org/10.1016/j.ijme.2023.100790

Lin, S., & Crosthwaite, P. (2024). The grass is not always greener: Teacher vs. GPT-assisted written corrective feedback. *System, 127*, 103529.

Luckin, R. (2017). Towards artificial intelligence-based assessment systems. *Nature Human Behaviour, 1*(3). https://doi.org/10.1038/s41562-016-0028

Martin, J. R., & White, P. R. R. (2005). *The language of evaluation: Appraisal in English*. Palgrave.

McGrath, L., & Kuteeva, M. (2012). Stance and engagement in pure mathematics research articles: Linking discourse features to disciplinary practices. *English for Specific Purposes, 31*(3), 161–173.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111*(2), 361–365. https://doi.org/10.1037/0033-2909.111.2.361

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook*. Sage.

Mizumoto, A., Yasuda, S., & Tamura, Y. (2024). Identifying ChatGPT-generated texts in EFL students' writing: Through comparative analysis of linguistic fingerprints. *Applied Corpus Linguistics, 4*(3), Article 100106. https://doi.org/10.1016/j.acorp.2024.100106

O'Donnell, M. (2008). *UAM corpus tool. Universidad Autónoma de Madrid. Retrieved from O'Donnell, M. (2008). UAM corpus tool*. Universidad Autónoma de Madrid. Retrieved from www.Wagsoft.com/CorpusTool.

Oh, P. S., & Lee, G. G. (2024). Confronting imminent challenges in humane epistemic agency in science education: An interview with ChatGPT. *Science & Education*, 1–27. https://doi.org/10.1007/s11191-024-00515-1

OpenAI. (2023). *How ChatGPT and our foundation models are developed | OpenAI Help Center*. Openai.com. https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed.

Polverini, G., & Gregorcic, B. (2024). How understanding large language models can inform the use of ChatGPT in physics education. *European Journal of Physics, 45*(2), Article 025701. https://doi.org/10.1088/1361-6404/ad1420

Qiu, X., & Jiang, F. (2021). Stance and engagement in 3MT presentations: How students communicate disciplinary knowledge to a wide audience. *Journal of English for Academic Purposes, 51*, Article 100976. https://doi.org/10.1016/j.jeap.2021.100976

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. *In proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000)* (pp. 1–6), 1-8 October 2000, Hong Kong.

Rowland, D. R. (2023). Two frameworks to guide discussions around levels of acceptable use of generative AI in student academic research and writing. *Journal of Academic Language and Learning, 17*(1), T31–T69.

Stahl, B. C., & Eke, D. (2024). The ethics of ChatGPT–Exploring the ethical issues of an emerging technology. *International Journal of Information Management, 74*, Article 102700.

Stokel-Walker, C. (2022). AI bot ChatGPT writes smart essays — should academics worry? *Nature*. https://doi.org/10.1038/d41586-022-04397-7

Swales, J. M. (1990). *Genre analysis*. Cambridge University Press.

Tardy, C. (2004). The role of English in scientific communication: Lingua franca or Tyrannosaurus rex? *Journal of English for Academic Purposes, 3*(3), 247–269. https://doi.org/10.1016/j.jeap.2003.10.001

Yang, S., Chen, S., Zhu, H., Lin, J., & Wang, X. (2024). A comparative study of thematic choices and thematic progression patterns in human-written and AI-generated texts. *System, 126*, Article 103494. https://doi.org/10.1016/j.system.2024.103494

Yeadon, W., Inyang, O.-O., Mizouri, A., Peach, A., & Testrow, C. P. (2023). The death of the short-form physics essay in the coming AI revolution. *Physics Education, 58* (3), Article 035027. https://doi.org/10.1088/1361-6552/acc5cf

Yeralan, S., & Lee, L. A. (2023). Generative AI: Challenges to higher education. *Sustainable Engineering and Innovation, 5*(2), 107–116. https://doi.org/10.37868/sei.v5i2.id196

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

Zhishan MO is a graduate student in applied linguistics at the University of Queensland. She is interested in corpus-based studies of academic writing as well as neurolinguistic accounts of second language acquisition.

Peter Crosthwaite is an Associate Professor in the School of Languages and Cultures at the University of Queensland. His areas of research and supervisory expertise include corpus linguistics and the use of corpora for language learning (known as 'data-driven learning'), as well as computer-assisted language learning, and English for General and Specific Academic Purposes. He is the Editor-in-Chief for the Australian Review of Applied Linguistics (from 2024) and serves on the editorial boards of the Q1 journals IRAL, Journal of Second Language Writing, Journal of English for Academic Purposes, and System.