Iperstoria
2025, no. 26
ISSN: 2281-4582

DOI: 10.13136/2281-4582/2025.i26.1705
iperstoria.it

Writing with a Reader in Mind

The Rhetorical Gap Between Genuine EFL Student Essays and LLM-Generated Essays

Marco Bagni
University of Modena and Reggio Emilia
ORCID: https://orcid.org/0000-0002-0785-9462
Email: marco.bagni@unimore.it

Keywords

EFL

LLMs

Academic writing

Metadiscourse

AI-generated texts

Abstract

Recent research has highlighted the potential of large language models (LLMs) such as ChatGPT to support the development of ESP writing skills. However, concerns have also emerged about learners delegating writing tasks to these tools, potentially undermining learners' motivation and autonomy. LLMs' capacity to generate human-like text challenges the ability to ensure the authenticity of EFL students' writing. Some studies have investigated LLM-generated text characteristics, comparing them with student writing. However, research on LLMs in EFL education remains limited. This paper seeks to address this gap and advance critical understanding of LLM-generated text. It reports on a small-scale study comparing three corpora: genuine undergraduate EFL student essays, suspected LLM-generated essays, and texts produced by ChatGPT 4.0 and DeepSeek-V3. The study examined patterns of interactive and interactional metadiscourse in the essays. Combining quantitative analysis with qualitative interpretation, it was found that although interactive features have limited explanatory power in distinguishing between LLM-generated text and genuine student essays, LLMs place strong emphasis on explicit cohesion, coherence and logical organization, favoring an objective and factual style. Most importantly, while they approximate human stance-taking, LLMs show very limited capacity to model audience and engage readers. This pattern was mirrored in the essays suspected to be LLM-generated. In contrast, genuine student writing was found to display a more personal tone and substantially greater reader engagement, regardless of target language proficiency.

1. Introduction

Recent surveys (e.g., Rong and Chun 2024) indicate ChatGPT's widespread popularity among students across educational levels. In higher educational settings, its rapid and widespread adoption has raised critical concerns regarding academic integrity, as students' reliance on the tool may undermine independent learning and compromise the authenticity of their work (Creely 2024; Kohnke et al. 2023; Moorhouse et al. 2023). In the context of English language teaching (ELT), significant concerns relate to the risk that learners may fully delegate writing tasks to the tool, potentially threatening their motivation and autonomy in writing skills development (Barrot 2023).

ChatGPT is a prominent example of a Generative AI (GenAI) application known as a Large Language Model (LLM), a type of machine learning model that understands and generates human-like language, operating across multiple languages, based on statistical patterns calculated in extensive training datasets (Cain 2024). The apparently human-like quality of LLM-generated text raises critical concerns regarding the authenticity of English as a foreign language (EFL) student writing, with further practical implications for assessment, as EFL teachers might struggle to identify genuine student writing (Alexander et al. 2023).

While some studies assessed the ability of humans and machines to detect AI-generated output and others looked at the characteristics of AI-generated texts (see section 2), research on LLMs in EFL education is still in its infancy. Further studies are required to contribute to improving critical literacy of such tools.

This paper aims to address this research gap. It reports on a small-scale comparative analysis of genuine and suspected LLM-generated essays from first-year Communication Sciences students' English exam at the University of Modena and Reggio Emilia, and essays generated by two prominent LLMs: ChatGPT 4.0 and DeepSeek-V3.

This study draws on Hyland's (2005a) model of metadiscourse, referring to linguistic resources used by writers to structure their texts, express their attitudes toward content and engage readers. Emphasizing the writing process over the final product, it aims to explore different patterns of metadiscourse in the essays and see whether distinguishing characteristics can be identified comparing LLM-generated text and genuine student writing, regardless of student academic proficiency.

This research is intended to offer practical insights to EFL teachers who may be confronting instances of academic misconduct involving LLM-generated texts, while also seeking to promote critical GenAI literacy.

2. LLM-generated text and human writing

Recent academic research (e.g., Cogo et al. 2024) attests to the significant impact that the sudden and widespread uptake of ChatGPT has had on English Language Teaching (ELT). A strand of research (Ramazani et al. 2025; Teng 2025; Su et al. 2024; Yoo-Jean 2024; Woo et al. 2024; Barrot 2023; Huang 2023; Warschauer et al. 2023) has discussed how ChatGPT, as well as other LLMs, can support writing skills development for EFL learners. Other studies have focused on EFL teachers' ability to distinguish genuine human writing from LLM-generated output (De Wilde 2024; Alexander et al. 2023), finding that teachers prevalently assess text source based on deficit criteria, typically regarding the absence of common learners' errors as proof of AI-generated output, while also revealing a tendency to attribute personal and creative texts to human writers. It has also been noted that existing AI-detection tools are limited in their ability to distinguish between human- and AI-generated texts. In particular, while they can often identify fully AI-generated output, they struggle to detect hybrid texts that combine human and machine-generated writing (Alexander et al. 2023).

Studies that investigated the characteristics of ChatGPT-generated output (Basic et al. 2023; Bishop 2023; Borji 2023; Frye 2023; Fyfe 2023), concluded that this tends to be excessively literal, neutral and superficial. Some studies (Jiang and Hyland 2025a; 2025b; Mo and Crosthwaite 2025) analyzed the differences in the use of metadiscourse (Hyland 2005a) between argumentative writing produced by LLMs and students in English L1 contexts. Findings have highlighted that human writing displays a broader and more varied use of interactional features of stance and particularly reader engagement. Conversely, LLMs exhibit high structural coherence but tend to lack human-like contextual understanding, limiting their ability to adapt rhetorically to specific audiences. It has been suggested that this limitation may stem from insufficient exposure to relevant, domain-specific training data (Jiang and Hyland 2025a, 2025b; Mo and Crosthwaite 2025).

These findings broadly accord with research works that compared EFL students' essays with ChatGPT-generated essays, finding that LLM-generated text lacks the nuanced authorship and contextual appropriateness characteristic of human writing (Amirjalili et al. 2024), and revealing greater use of epistemic markers conveying personal stance in students' writing (Herbold et al. 2023). However, while a few studies (see e.g. Yoon 2021; Ho and Li 2018; Zhao 2017) have variously examined metadiscourse (Hyland 2005a; 2005b) in EFL argumentative writing, there is a paucity of research providing textual evidence of the differences in patterns of metadiscourse between AI-generated and EFL student writing across various genres.

The present investigation extends analyses of metadiscursive differences between AI-generated and human writing into an EFL context with a focus on three essay types: compare-and-contrast, argument, and discussion. Although teachers may intuitively recognize writing that is not genuinely produced by their students, examining differences in the use of metadiscourse may provide insights into both LLMs' text generation capabilities and limitations, as well as into the creative process of human writing.

3. Metadiscourse: writing as interaction

This study approaches writing as a socially situated act of interaction and focuses on the ways essay authors employ linguistic resources to manage the discourse, project an authorial presence, and engage the intended reader. It draws on Hyland's (2005a) framework of metadiscourse, distinguishing:

1. The interactive dimension, concerning "ways of organizing discourse, rather than experience" (Hyland 2005a, 49) in the text, including:
   - code glosses, supplying "additional information, by rephrasing, explaining or elaborating what has been said" (Hyland 2005a, 51)
   - endophoric markers, referring to parts of the text and guiding readers through the discourse
   - evidentials, guiding "the reader's interpretation" and establishing "an authorial command of the subject"
   - frame markers, ordering arguments (and not events of facts in time)
2. Transition markers helping the audience interpret logical connections in arguments the interactional dimension, conveying authorial stance and engaging the reader. Stance is "the writer's expression of a textual 'voice'" (Hyland 2005a, 49) and is realized in texts through the following resources:
   - hedges, reflecting the writer's decision to acknowledge alternative viewpoints and withhold full commitment to the claims
   - boosters, allowing writers to express certainty in their claims and reinforcing a sense of alignment with the audience
   - attitude markers, conveying the writers' emotional or evaluative stance toward the propositional content
   - self-mentions (i.e., first-person pronouns and possessive adjectives) explicitly signalling the writer's presence. Engagement refers to the resources used by writers to include readers as participants and direct them toward intended interpretations, comprising:
   - reader mentions (e.g., the pronouns *you*, *your* and the inclusive *we*) pulling readers into the text
   - directives, instructing the reader to take specific action or adopt a perspective aligned with the writer's intentions
   - appeals to shared knowledge
   - personal asides, providing a brief comment or reflection on preceding content
   - questions encouraging reader involvement

4. Method

4.1 Data collection

Initially, 74 timed essays were collected. They had been written in pen and paper by non-English majors in their first year of the BA in Communication Sciences at the University of Modena and Reggio Emilia, for the in-class final exam of the mandatory English language course taught by the author of this paper. The course included a module on reading and writing skills, with particular emphasis on essay writing across three sub-genres: compare-and-contrast, argument, and discussion – with the proficiency benchmark set at B1+. In line with this benchmark, the teacher had instructed students to prioritize coherence and cohesion and to re-elaborate course content in a personal way, without strictly adhering to academic style conventions typically expected of highly proficient students and scholars – such as the use of impersonal language – though these conventions were illustrated in textbook extracts included in the course materials, which the students had been explicitly told to disregard.

Two exam sheet versions had been created, each including a title for a distinct essay type: compare and contrast, argument, and discussion, as summarized in Table 1. The students had been asked to choose one title and write approximately 200 words.

Exam sheet 1	
A Compare and contrast	*Learning English in the past (before the advent of the Internet) versus learning English nowadays, in the era of the Internet.*
B Argument	*English language teachers should base learning activities on more varieties of English, not only on one Standard variety.*
C Discussion	*The global spread of English is a threat to the world's linguistic and cultural diversity.*

Exam sheet 2	
A Compare and contrast	*Learning English as a foreign language in the past – before the advent of the internet – versus learning English today, in the era of the Internet.*
B Argument	*The use of English as a lingua franca makes communication more efficient, in settings that involve speakers of mutually unintelligible languages (such as, for instance, Italian, German, Chinese, Korean, etc.).*
C Discussion	*English learning cannot be based exclusively on Standard English.*

Tab. 1: Essay titles

During the exam grading process, the teacher identified 16 essays that bore little to no connection with the course content and, most notably, contained misspellings and conspicuous errors embedded within otherwise fluent and polished academic English. These inconsistencies led the teacher to suspect that the texts had been either inattentively copied from an uncredited GenAI tool or partially copied and subsequently modified.

All essays were transcribed verbatim, preserving misspellings and morphosyntactic errors, using the Microsoft Word Dictate tool. The accuracy of the transcription was manually checked. The suspect essays were set apart, and two distinct corpora were compiled using Sketchengine, a web-based software providing a comprehensive set of corpus-analysis tools. One corpus, henceforth referred to as GSE, included 58 essays genuinely written by students; the other, henceforth referred to as SSE, included the 16 suspect essays.

Subsequently, a third corpus, henceforth referred to as AGE, was compiled with 12 essays generated by the free browser versions of ChatGPT 4.0 and DeepSeek-V3. The decision to use two LLMs was driven by the need to introduce diversity and enhance reliability. Each LLM was prompted to produce one essay for each title included in the exam sheets (see Table 1) with zero-shot prompts (Schulhoff et al. 2024). Zero-shot prompts depend on the LLM's internal knowledge, with the expectation that it will generate a response using only the information contained in the prompt and without additional contextual information and examples. The prompts included only the essay title and the indication of approximately 200 words as the required essay length. This approach was suggested by the in-class exam context, based on the assumption that if students had used GenAI to cheat, they had done so very discreetly and therefore they had very likely input only the essay title as a prompt, without any additional instructions. Finally, within each corpus, three sub-corpora were compiled, each corresponding to a distinct essay type, as summarized in Table 2.

Corpus	Subcorpus	Essay type	# of Docs	Token count
AGE	*A*	Compare and contrast	*4*	895
AGE	*B*	Argument	*4*	846
AGE	*C*	Discussion	*4*	937
AGE			***12***	**2678**
SSE	*A*	Compare and contrast	*7*	1463
SSE	*B*	Argument	*3*	471
SSE	*C*	Discussion	*6*	992
SSE			***16***	**2926**
GSE	*A*	Compare and contrast	*20*	4592
GSE	*B*	Argument	*20*	4666
GSE	*C*	Discussion	*18*	4174
GSE			***58***	**13432**

Tab. 2: Corpora details

4.2 Data analysis

The present research employed a mixed-method approach, combining quantitative automated frequency analysis and qualitative, interpretative analysis to ensure comprehensiveness and accuracy in metadiscourse items identification.

For each corpus and respective subcorpora, a wordlist was created using the dedicated Sketchengine tool 'Wordlist'. The wordlists showed all the words' absolute frequency, relative frequency, document frequency and relative document frequency. Subsequently, sets of interactive and interactional metadiscourse items were compiled based on Hyland's (2005a;

2005b) framework. These compilations were then organized into two separate CSV files – one containing interactive and the other interactional resources.

Following this, ChatGPT 4.0 Plus was instructed to search the corpora for each item included in the lists. Automated search enabled efficient and quick retrieval of metadiscourse items from the corpora; however, ChatGPT was not employed to assess the function of each item, as it had previously demonstrated unreliability in this task, with several inconsistencies and omissions in its output. Therefore, the results obtained from the ChatGPT-powered item search were manually checked using the Sketchengine 'Concordance' tool, to establish whether the features identified performed interactive or interactional meanings in the texts.

The subsequent phase involved compiling the results into two Excel files – one for interactive and one for interactional metadiscourse – organized as itemized lists detailing each word's absolute and relative frequency, document frequency and relative document frequency.

Finally, in each file, the list of items was then integrated with items emerging from close reading of the students' essays. Metadiscourse is "realized by an open-ended set of language items" (Hyland 2005a, 37) and, most importantly, several students' essays displayed forms that resulted from L1 transfer and L2 proficiency. These forms were carefully analyzed to understand the students' communicative purpose, and it was concluded that some of them revealed the intention to realize interactive or interactional meanings. For instance, a student (1) used a loan translation of an Italian expression corresponding to *it is necessary to say / it must be said*, which was counted among the directives (engagement).

(1) To conclude, it is to say that it is hard to predict what the future will bring

(GSE A essay)

The absolute and relative frequency, document frequency, and relative document frequency of these items were manually calculated, and a final version of the Excel files was compiled to reflect the comprehensive search results.

5. Results

As shown in Table 3, the item search identified nearly twice as many metadiscourse features per million words (relative frequency) in genuine student essays compared to both LLM-generated and suspect student essays, with a remarkable gap in interactional elements.

	Frequency			Rel. Freq.			Document freq.			Rel. Doc. freq. (%)		
	AGE	SSE	GSE	AGE	SSE	GSE	AGE	SSE	GSE	AGE	SSE	GSE
INTERACTIVE	70	68	425	26138.91	23239.92	31640.86	12	15	58	100%	93.75%	100%
INTERACTIONAL	51	53	701	19044.06	18113.47	50297.95	12	15	58	100%	93.75%	100%
Tot.	**121**	**121**	**1126**	**45182.97**	**41353.39**	**81938.81**	**12**	**16**	**58**	**100%**	**100%**	**100%**

Tab. 3: Metadiscourse items in the three corpora

In the following sections, interactive and interactional features are analyzed first by comparing the three main corpora, and then the distinct subcorpora within the main ones.

5.1 Interactive features

The search for interactive metadiscourse features revealed no substantial variation across the three main corpora (Table 4).

Interactive items	**Frequency**			**Relative freq.**		
	AGE	SSE	GSE	AGE	SSE	GSE
Code glosses	21	14	101	7841.67	4784.69	7519.36
Endophoric markers	0	1	2	0	341.76	148.90
Evidentials	3	2	8	1120.24	683.53	595.59
Frame markers	17	10	83	6348.02	3417.64	6179.27
Transitions	29	41	231	10828.98	14012.30	17197.74
tot.	**70**	**68**	**425**	**26138.91**	**23239.92**	**31640.86**

	Document freq.			**Rel. Doc. freq.**		
	AGE	SSE	GSE	AGE	SSE	GSE
Code glosses	9	9	47	75.00%	56.25%	81.03%
Endophoric markers	0	1	2	0.00%	6.25%	3.45%
Evidentials	3	2	5	25.00%	12.50%	8.62%
Frame markers	10	7	49	83.33%	43.75%	84.48%
Transitions	9	12	51	75.00%	75.00%	87.93%
tot.	**12**	**15**	**58**	**100.00%**	**93.75%**	**100.00%**

Tab. 4: Summary of interactive resources in the three main corpora

However, SSE exhibited a notably lower frequency of code glosses, particularly in argumentative essays (see 5.1.2). A plausible explanation for this may be that the students who engaged in academic misconduct omitted code glosses – possibly present in the original LLM-generated source – not necessarily because they deemed them unimportant, but more likely as a time-saving strategy, given the time-pressured and monitored setting of the in-class exam. Given the short length of the essays, the almost complete absence of endophoric markers was to be expected, and the significance of those found in SSE (see also 5.1.1) and GSE mainly resided in the fact that they coincided with self-mentions – a key component of interactional metadiscourse (see 5.2).

AGE essays featured a slightly higher relative frequency of evidentials, although due to the small sample size, the findings should be interpreted with caution and cannot support definitive conclusions. Most notably, AGE evidentials – found once in the argument essays and twice in the discussion essays – were used within the thematic progression to introduce arguments and counterarguments, although they typically referred to generic subjects, as exemplified in (2):

(2) Nevertheless, some argue that bilingualism allows people to embrace English while preserving their mother tongues (AGE C essay)

The only instances of attribution to identifiable sources drawn from course content appeared exclusively in the GSE corpus (see 5.1.1 and 5.1.2).

SSE also displayed a markedly lower use of frame markers, particularly in compare-and-contrast essays (SSE A, see 5.1.1), and secondarily in discussion essays (SSE C, see 5.1.3), whereas their relative frequency in the argument essays (SSE B) exceeded that found in both AGE and GSE essays of the same genre (see 5.1.2).

The comparatively lower frequency of transitions in AGE reflected the LLMs' tendency to present information concisely and directly.

5.1.1 Compare-and-contrast essays

Table 5 reports the distribution of interactive features in the compare-and-contrast essays.

Interactive items	**Frequency**			**Relative freq.**		
	AGE A	SSE A	GSE A	AGE A	SSE A	GSE A
Code glosses	3	8	37	3351.96	5468.22	4057.49
Endophoric markers	0	1	1	0.00	683.53	217.77
Evidentials	0	0	0	0.00	0.00	0.00
Frame markers	4	2	24	4469.27	1367.05	5226.48
Transitions	5	3	25	5586.59	3351.96	5444.25
tot.	**12**	**14**	**87**	**13407.82**	**10870.75**	**14945.99**
	Document freq.			**Rel. Doc. freq.**		
	AGE A	SSE A	GSE A	AGE A	SSE A	GSE A
Code glosses	3	5	17	75.00%	71.43%	85.00%
Endophoric markers	0	1	1	0.00%	14.29%	5.00%
Evidentials	0	0	0	0.00%	0.00%	0.00%
Frame markers	4	2	17	100.00%	28.57%	85.00%
Transitions	2	3	13	50.00%	42.86%	65.00%
tot.	**4**	**6**	**20**	**100.00%**	**85.71%**	**100.00%**

Tab. 5: Interactive resources in the compare-and-contrast essays

The nature of the compare-and-contrast essay – a genre that aims to show similarities and differences between two situations – might account for the relatively infrequent use of code glosses in AGE A and GSE A. While the frequency of code glosses in SSE A was higher than in AGE A and GSE A, relative document frequency values across the three subcorpora were broadly comparable. Also, most code glosses in SSE A essays (5 out of 8) consisted of the phrase *such as*, used to introduce similar examples of teaching materials, methods, and resources, as illustrated in (3):

(3) Access to authentic English materials, such as newspapers, radio broadcasts, was limited and expensive. (SSE A essay)

This pattern suggested a possible shared GenAI source for those five essays, which also displayed notable thematic similarities. Notwithstanding this observation, the limited size of the corpus precludes any definitive conclusions regarding the discrepancy in code gloss frequency across the subcorpora.

The endophoric marker, which coincided with a self-mention (see 5.2), found in 1 SSE A essay, suggested that an element of personal expression had been introduced by the student into the otherwise neutral tone of the AI-generated text:

(4) Historically, English learning involved primarily face to face interaction and this created a personal learning environment, but it limited exposure to native speakers and real-world language use. In contrast, the Internet has revolutionized interaction and communication in English learning. As I said online platforms allow learners to connect with native speakers globally, enhancing their speaking and listening skills. (SSE A essay)

The absence of frame markers in 5 out of the 7 SSE A essays may be attributed to the inherent characteristics of LLM-generated output in response to the students' prompts, as evidenced by essays structured as simple lists of compare-and-contrast points. Although the relative frequency of frame markers in AGE A was higher than in SSE A, the AGE A essays similarly consisted of extensive yet concise lists of topics and, notably, included only a single frame marker each – the phrase *in conclusion* typically used to introduce a concluding statement. Notwithstanding these considerations, the size of both AGE A and SSE A limits the extent to which definitive conclusions can be drawn.

Transition markers were used more sparsely across all three subcorpora compared to the main corpora, a pattern that may also be attributed to genre, with compare-and-contrast essays encouraging a more linear mode of exposition and potentially requiring fewer explicit logical connectors than argumentative or discussion essays.

5.1.2 Argument essays

Table 6 summarizes the search results for interactive items in the argument essays.

Interactive items	Frequency			Relative freq.		
	AGE B	SSE B	GSE B	AGE B	SSE B	GSE B
Code glosses	14	1	37	16607.35	2123.14	7929.70
Endophoric markers	0	0	0	0	0	0
Evidentials	1	0	3	1186.24	0	642.95
Frame markers	6	5	33	7117.44	10615.71	7072.44
Transitions	16	9	113	18979.83	19108.28	28504.07
tot.	**37**	**15**	**186**	**43890.87**	**31847.13**	**44149.16**
	Document freq.			**Rel. Doc. freq.**		
	AGE B	SSE B	GSE B	AGE B	SSE B	GSE B
Code glosses	4	1	17	100.00%	33.33%	85.00%
Endophoric markers	0	0	0	0.00%	0.00%	0.00%
Evidentials	1	0	2	25.00%	0.00%	10.00%
Frame markers	2	2	18	50%	66.67%	90.00%
Transitions	4	3	20	100.00%	100.00%	100.00%
tot.	**4**	**3**	**20**	**100.00%**	**100.00%**	**100.00%**

Tab. 6: Interactive resources in the argument essays

While the total frequency values of interactive devices across the three subcorpora did not reveal any substantial differences, a notably higher occurrence of code glosses was observed in AGE B. As illustrated in Table 7, these included features such as the phrase *such as*, the abbreviation *e.g.*, parentheses, and dashes used to introduce examples and explanations – elements characteristic of academic register and indicative of the concise style typical of AGE texts. Although code glosses were comparatively less frequent in GSE B, their use was more diverse, notably including examples, suggesting not only a lower degree of familiarity with academic conventions but also a greater degree of textual personalization by the students.

	AGE B				**GSE B**				**SSE B**			
Item	**Freq.**	**Rel. freq.**	**Doc. Freq.**	**Rel. Doc. Freq.**	**Freq.**	**Rel. freq.**	**Doc. Freq.**	**Rel. Doc. Freq.**	**Freq.**	**Rel. freq.**	**Doc. Freq.**	**Rel. Doc. Freq.**
(	3	3559	1	25%	0	0	0	0%	8	1715	6	30%
–	2	2372	2	50%	0	0	0	0%	0	0	0	0%
e.g.	2	2372	1	25%	0	0	0	0%	1	214	1	5%
for instance	1	1186	1	25%	0	0	0	0%	9	1929	8	40%
for example	0	0	0	0%	0	0	0	0%	6	1286	6	30%
such as	4	4745	3	75%	1	2123	1	33%	6	1286	6	30%
like	2	2372	2	50%	0	0	0	0%	4	857	4	20%
in fact	0	0	0	0%	0	0	0	0%	1	214	1	5%
called	0	0	0	0%	0	0	0	0%	1	214	1	5%
which means	0	0	0	0%	0	0	0	0%	1	214	1	5%

Tab. 7: Code glosses in argument essays

As noted earlier (see 5.1), the presence of only one code gloss in SSE B may be attributed to students who engaged in academic misconduct, omitting such elements – likely as a practical, time-saving measure in the time-constrained and proctored exam setting.

Interestingly, GSE B featured two attributions to cited sources (evidentials), used by one student (5) who correctly supported her argument by referring to sources cited in the course content material.

(5) According to McLuhan the world works today as a "global village" (…) according to statista statistics, (GSE B essay)

The other evidential found in GSE B instead of a rather vague nature, as exemplified by (6):

(6) It has been argued that the uneven level of fluency of NNES causes a loss of nuance (GSE B essay)

As previously observed (see 5.1), the single evidential found in AGE B was also referred to a generic subject:

(7) However, some argue that English dominance disadvantages non-native speakers. (AGE B essay)

Although AGE B and GSE B displayed nearly identical relative frequency values for frame markers, and SSE B showed a comparatively higher frequency value for such items, the relative document frequency values indicated a more widespread use of frame markers in genuine student essays. Given the limited corpus size, it is difficult to establish a reliable correlation between the presence of a human touch and the use of frame markers. Nevertheless, it is worth observing that AGE B and SSE B only included text sequencing and text stages labelling devices (*first*, *firstly*, *secondly*, *in conclusion*), whereas devices used to announce goals and shift topics were found exclusively in GSE B. Notably, the expressions used to announce goals co-occurred with self-mentions (see 5.2), as exemplified by (8) and (9):

(8) the use of English as a lingua franca makes the communication more efficient and I am going to explain my point of view (GSE B essay)

(9) First of all, we are going to talk about those who think that the answer to this question is yes (GSE B essay)

Argument essays overall also featured a substantially higher number of transitions than compare-and-contrast essays, once again suggesting genre constraints. Moreover, GSE B featured a markedly more frequent and also more varied use of transition markers than SSE B and AGE B, indicating distinctions in how arguments were structured between LLM-generated text and student writing. Table 8 presents the comprehensive results of the search for transition signals in argument essays.

	AGE B				GSE B				SSE B			
Item	Freq.	Rel. freq.	Doc. Freq.	Rel. Doc. Freq.	Freq.	Rel. freq.	Doc. Freq.	Rel. Doc. Freq.	Freq.	Rel. freq.	Doc. Freq.	Rel. Doc. Freq.
additionally	0	0	0	0%	1	2123	1	33%	0	0	0	0%
also	1	1186	1	25%	1	2123	1	33%	7	5787	5	25%
and	4	4745	3	75%	1	2123	1	33%	27	5787	12	60%
because	0	0	0	0%	0	0	0	0%	24	5144	15	75%
besides	0	0	0	0%	0	0	0	0%	1	214	1	5%
but	1	1186	1	25%	1	2123	1	33%	23	4929	15	75%
furthermore	1	1186	1	25%	1	2123	1	33%	1	214	1	5%
however	2	2372	2	50%	1	2123	1	33%	4	857	3	15%
moreover	1	1186	1	25%	0	0	0	0%	1	214	1	5%
nevertheless	0	0	0	0%	0	0	0	0%	3	643	3	15%
rather	2	2372	1	25%	1	2123	1	33%	0	0	0	0%
since	0	0	0	0%	0	0	0	0%	1	214	1	5%
so	0	0	0	0%	0	0	0	0%	13	2786	8	40%
although	0	0	0	0%	0	0	0	0%	5	1072	5	25%
even though	0	0	0	0%	0	0	0	0%	2	429	2	10%
though	0	0	0	0%	1	2123	1	33%	0	0	0	0%
thus	1	1186	1	25%	0	0	0	0%	0	0	0	0%
while	3	3559	3	75%	1	2123	1	33%	1	214	1	0%

Tab. 8: Transition signals in argument essays

It can be seen that GSE B exhibited a broader range of logical relations compared to AGE B and SSE B. These included not only contrast and concession (*but*, *however*, *rather*, *though*, *while*) but also cause/reason (e.g., *because*) and effect/result (e.g., *so*). Notably, students who employed causal connectors often did so in relation to classroom topics, drawing on their knowledge of course content to support their arguments.

5.1.3 Discussion essays

Table 9 summarizes the search results for interactive items in the discussion essays.

Interactive items	Frequency			Relative freq.		
	AGE C	SSE C	GSE C	AGE C	SSE C	GSE C
Code glosses	4	5	27	4268.94	5040.32	6468.62
Endophoric markers	0	0	1	0.00	0.00	239.58
Evidentials	2	2	5	2134.47	2016.13	1197.89
Frame markers	7	3	26	7470.65	3024.19	6229.04
Transitions	8	29	93	8537.89	28225.81	22280.79
tot.	**21**	**39**	**152**	**22411.95**	**38306.45**	**36415.91**
	Document freq.			**Rel. Doc. freq.**		
	AGE C	SSE C	GSE C	AGE C	SSE C	GSE C
Code glosses	2	3	13	50.00%	50.00%	72.22%
Endophoric markers	0	0	1	0.00%	0.00%	5.56%
Evidentials	2	2	3	50.00%	33.33%	16.67%
Frame markers	4	3	14	100.00%	50.00%	77.78%
Transitions	3	6	18	75.00%	100.00%	100.00%
tot.	**4**	**6**	**18**	**100.00%**	**100.00%**	**100.00%**

Tab. 9: Interactive resources in the discussion essays

Although minimal variation was found across the three subcorpora, the slightly higher frequency of code glosses in GSE C may suggest greater emphasis on explanation or reader guidance in essays genuinely written by students as opposed to those generated by LLMs.

All the evidentials found in the discussion essays subcorpora did not refer to any specified source but only to generic subjects:

(10) Nevertheless, some argue that bilingualism allows people to embrace English while preserving their mother tongues (AGE C essay)

(11) However, some scholars argue that English should not only be seen as a threat, but also as an asset (SSE C essay)

(12) Some people believe that the spread of English is very useful (GSE C essay)

The relatively sparser use of frame markers in SSE C, compared to both AGE C and GSE C, may be partly attributed to the hybrid nature of the essays, some of which were particularly short and appeared to be partially copied from more elaborate sources. Additionally, it is possible that the GenAI tool used to generate other SSE C essays produced limited discourse framing simply as a result of its statistically driven output. However, once again, given the limited size of the corpus and the indirect nature of the evidence, no firm conclusions can be drawn.

Interestingly, even more so than AGE B essays, AGE C essays exhibited a preference for a factual mode of presentation and a pronounced tendency toward conciseness, as evidenced by the substantially reduced use of transition markers. The high frequency of transitions in SSE C stood out. However, most transition markers in this subcorpus (14 out of 29) occurred in just 2 essays, each containing 7 such items, which appeared unmistakably hybrid. These essays addressed the discussion essay title in Exam sheet 2, on the global spread of English threatening the world's linguistic and cultural diversity (see Table 1), and advanced arguments also found in the other AGE C and SSE C essays from the same exam sheet. Significantly, these arguments – and especially the lexical choices through which they were expressed (e.g., *the decline, extinction* or *loss of minority languages*, *the marginalization of local languages and cultures*) – were absent from the course materials (which addressed the title topic from a rather different perspective). This strongly suggested that the authors of these SSE C essays had re-elaborated uncredited AI-generated output, restructuring sentences and inserting transitions to adapt the text. Rather than casting doubt on the non-genuine nature of the essays, the high frequency of transitions in SSE C thus provided evidence in support of it.

5.2 Interactional features

The interactional items search results revealed a clear disparity across the three corpora (Table 10). GSE displayed a relatively more frequent use of stance markers and a significantly more consistent use of engagement features than both AGE and SSE, which revealed instead largely similar patterns.

Interactional items	**Frequency**			**Relative freq.**		
	AGE	SSE	GSE	AGE	SSE	GSE
Hedges	29	26	98	10828.98	8885.85	7296.00
Boosters	5	3	78	1867.06	1025.29	5807.03
Attitude	10	9	86	3734.13	3075.87	6402.62
Self-mentions	3	10	79	1120.24	3417.64	5881.48
STANCE TOT.	**47**	**48**	**341**	**17550.41**	**16404.65**	**25387.13**
Reader mentions	0	2	279	0.00	683.53	20771.29
Directives	2	2	55	746.83	683.53	4094.70
Appeals to shared knowledge	1	0	4	373.41	0.00	6.90
Asides	0	1	7	0.00	341.76	12.07
Questions	1	0	15	373.41	0.00	25.86
ENGAGEMENT TOT.	**4**	**5**	**360**	**1493.65**	**1708.82**	**24910.82**
	Document freq.			**Rel. Doc. freq.**		
	AGE	SSE	GSE	AGE	SSE	GSE
Hedges	12	13	37	100.00%	81.25%	63.79%
Boosters	5	2	36	41.67%	12.50%	62.07%
Attitude	6	6	40	50.00%	37.50%	68.97%
Self-mentions	3	5	34	25.00%	31.25%	58.62%
STANCE TOT.	**12**	**15**	**58**	**100.00%**	**93.75%**	**100.00%**
Reader mentions	0	2	46	0.00%	12.50%	79.31%
Directives	2	2	24	16.67%	12.50%	41.38%
Appeals to shared knowledge	1	0	3	8.33%	0.00%	5.17%
Asides	0	1	7	0.00%	6.25%	12.07%
Questions	1	0	12	8.33%	0.00%	20.69%
ENGAGEMENT TOT.	**4**	**4**	**51**	**33.33%**	**25.00%**	**87.93%**

Tab. 10: Summary of interactional resources in the three main corpora

Comparing patterns of stance markers, one can see that, unlike GSE, AGE predominantly featured hedges, indicating that LLMs tended to present information in a highly impersonal manner, with minimal authorial presence, limited commitment to claims, and reduced affective engagement. The observed tendency may be partially attributed to the zero-shot prompt strategy employed, leading the LLMs to adopt a more detached stance. Stance patterns in SSE further suggested the hybrid nature of the essays: the relative frequency of hedges was higher than in GSE but lower than in AGE, while the relative frequency of self-mentions exceeded that of AGE yet remained below the levels observed in GSE.

Crucially, the search results clearly revealed that the distinguishing characteristic of genuine student essays did not lie so much in how students projected their persona (stance), but in how they involved the reader in the text (engagement), with an overwhelming presence of reader mentions in GSE essays. Nevertheless, it must also be observed that not all essays (79.31%) featured reader mentions. As Jiang and Hyland (2025a) observed, learners may exhibit reluctance to engage readers in a direct and personalized manner, possibly perceiving this strategy as characteristic of more intimate or informal registers. Although the teacher had advised the students to disregard textbook prescriptions discouraging personal expression (see 4.1), some may have avoided reader mentions in an effort to conform to perceived academic conventions.

Notwithstanding that possibility, this rhetorical strategy was notably absent in AGE. This tendency very likely resulted from the LLMs' training data patterns, which tend to neutralize any specific audience, resulting in output that lacks targeted rhetorical adaptation (Jiang and Hyland 2025a; 2025b).

The analysis of interactional markers distribution across the three essay types revealed more nuanced patterns.

5.2.1 Compare-and-contrast essays

The distribution of interactional markers in the three subcorpora of compare-and-contrast essays is summarized in Table 11.

Interactional items		**Frequency**			**Relative freq.**	
	AGE A	SSE A	GSE A	AGE A	SSE A	GSE A
Hedges	10	10	11	11173.18	6835.27	2395.47
Boosters	1	0	11	1117.32	0.00	2395.47
Attitude	0	0	7	0.00	0.00	1524.39
Self-mentions	0	7	23	0.00	4784.69	5008.71
STANCE TOT.	**11**	**17**	**52**	**12290.50**	**11619.96**	**11324.04**
Reader mentions	0	1	132	0.00	683.53	28745.64
Directives	0	0	3	0.00	0.00	653.31
Appeals to shared knowledge	0	0	0	0.00	0.00	0.00
Asides	0	1	2	0.00	683.53	435.54
Questions	0	0	4	0.00	0.00	871.08
ENGAGEMENT TOT.	**0**	**2**	**141**	**0.00**	**1367.05**	**30705.57**
		Document freq.			**Rel. Doc. freq.**	
	AGE A	SSE A	GSE A	AGE A	SSE A	GSE A
Hedges	4	6	8	100.00%	85.71%	40.00%
Boosters	1	0	7	25.00%	0.00%	35.00%
Attitude	0	0	6	0.00%	0.00%	30.00%
Self-mentions	0	2	8	0.00%	28.57%	40.00%
STANCE TOT.	**4**	**6**	**20**	**100.00%**	**85.71%**	**100.00%**
Reader mentions	0	1	17	0.00%	14.29%	85.00%
Directives	0	0	2	0.00%	0%	10.00%
Appeals to shared knowledge	0	0	2	0.00%	0%	10.00%
Asides	0	1	2	0.00%	14.29%	10.00%
Questions	0	0	0	0.00%	0%	0.00%
ENGAGEMENT TOT.	**0**	**1**	**17**	**0.00%**	**14.29%**	**85.00%**

Tab. 11: Interactional resources in the compare-and-contrast essays

The almost identical relative frequency of overall stance features across the three subcorpora, and most notably their sparser use in comparison to the main corpora, may be attributed to genre, with the compare-and-contrast essay arguably encouraging a more factual and objective mode of presentation. This hypothesis appears to be confirmed by the predominance of hedges and the complete absence of self-mentions and attitude markers in AGE A – which otherwise mirrored AGE in terms of stance features. Only the frequency value of self-mentions in GSE A approximated that of the main corpus (GSE). Some self-mentions overlapped with attitude markers, as exemplified in (13):

(13) Before Internet learning English was, in my opinion more difficult (GSE A essay)

Self-mentions in GSE A accounted for almost half of all stance items in the subcorpus, suggesting that some students felt it important to project their authorial persona even within a genre that arguably called for a more objective than subjective approach. The presence of self-mentions in 2 SSE A essays suggested instead that the students, possibly to evade detection, made a deliberate effort to infuse a personal touch into an otherwise impersonal text, as in (14):

(14) I feel that with today's resources learning English is much easier and more interactive, achieving fluency is faster and more enjoyable than before. Now I can watch English videos and films join online courses and even talk to native speaker through apps. (SSE essay)

The complete absence of engagement features in AGE A reinforced the observed tendency of LLMs to present propositional content in an impersonal and matter-of-fact manner. AGE A essays consisted largely of concise topic listings, lacking the rhetorical strategies needed to establish a relationship between writer, audience, and subject matter.

This characteristic of LLM-generated texts sharply contrasted with the extensive use of reader mentions in GSE A, representing the vast majority (132 out of 141 in 17 out of 20 essays) of engagement markers in the subcorpus. This feature indicated that several students aimed to create a conversational tone, textually positioning themselves and the teacher as participants sharing a common understanding and goals, as exemplified in (15) and (16):

(15) We certainly have more opportunities, but we shouldn't forget that the best way to learn a new language is to travel the world and talk with people. (GSE A essay)

(16) before Internet English was only learned in school or "study books" and if you want it to maybe listen someone talking in English you had to go to school or move in another country where English was a native Language (GSE A essay)

Interestingly, while accounting for only 4 out of 141 engagement markers, two GSE A essays also included questions as a strategy of dialogic involvement. Three of them (17) were used by a student whose essay stood out for being notably geared toward fostering immediacy and actively engaging the reader:

(17) and this is how slang was born but how is slang any different from regular English? Why do we decide to use slang words? How is learning English different now from how it was used then? (GSE A essay)

SSE A mirrored AGE A in its lack of engagement features, notwithstanding the presence of 1 reader mention and 1 aside in one essay, giving further evidence of the student's apparent intent to personalize the AI-generated text and, potentially, to evade detection.

5.2.2 Argument essays

Table 12 summarizes the stance and engagement features found in the three argument essays subcorpora.

Interactional items	**Frequency**			**Relative freq.**		
	AGE B	SSE B	GSE B	AGE B	SSE B	GSE B
Hedges	5	10	53	5931.20	21231.42	11358.77
Boosters	2	0	41	2372.48	0.00	8786.97
Attitude	5	5	34	5931.20	10615.71	7286.76
Self-mentions	3	3	44	3558.72	6369.43	9429.92
STANCE TOT.	**15**	**18**	**172**	**17793.59**	**38216.56**	**36862.41**
Reader mentions	0	0	51	0.00	0.00	10930.13
Directives	2	2	31	2372.48	4246.28	6643.81
Appeals to shared knowledge	0	0	3	0.00	0.00	642.95
Asides	0	0	4	0.00	0.00	857.27
Questions	0	0	5	0.00	0.00	1071.58
ENGAGEMENT TOT.	**2**	**2**	**94**	**2372.48**	**4246.28**	**20145.74**
	Document freq.			**Rel. Doc. freq.**		
	AGE B	SSE B	GSE B	AGE B	SSE B	GSE B
Hedges	4	3	18	100.00%	100.00%	90.00%
Boosters	2	0	17	50.00%	0.00%	85.00%
Attitude	3	3	17	75.00%	100.00%	85.00%
Self-mentions	3	3	15	75.00%	100.00%	75.00%
STANCE TOT.	**4**	**3**	**20**	**100.00%**	**100.00%**	**100.00%**
Reader mentions	0	0	12	0.00%	0.00%	60.00%
Directives	2	2	13	50.00%	66.67%	65.00%
Appeals to shared knowledge	0	0	2	0.00%	0.00%	10.00%
Asides	0	0	4	0.00%	0.00%	20.00%
Questions	0	0	5	0.00%	0.00%	25.00%
ENGAGEMENT TOT.	**2**	**2**	**17**	**50.00%**	**66.67%**	**85.00%**

Tab. 12: Interactional resources in the argument essays

While the overall stance patterns did not differ much from those observed in the compare-and-contrast essays, in comparison, both LLM-generated and genuine student essays of the argumentative type exhibited a few differences. Within AGE B, 3 essays featured each 1 self-mention, corresponding to an expression of agreement with the essay title statements (see Table 1), as exemplified in (18):

(18) I agree that the use of English as a lingua franca makes communication more efficient (AGE B essay)

In addition, AGE B featured a more limited use of hedges compared to AGE A (and AGE overall), suggesting that agreement with the title statement had prompted the LLMs to reduce their use of hedges, even though the limited occurrence of boosters also suggested a lack of emphasis in presenting propositional content. This tendency of LLMs to maintain a detached stance toward the topic was also reflected in the complete absence of boosters and a relatively higher frequency of hedges in SSE B. Stance features were instead significantly more frequent in GSE B than in GSE A, a pattern likely shaped by the nature of the argument essay. This essay genre arguably promotes greater authorial involvement with the topic, leading to stronger commitments to the reliability of the content, heightened affective stance, and a more pronounced authorial presence in the text.

The patterns of engagement features in GSE B broadly reflected the general pattern observed in GSE. However, reader mentions were much less frequent in the subcorpus of argument essays, suggesting that, in this case, too, some students may have deliberately refrained from engaging the reader directly to align with perceived academic conventions and avoid a tone that might have been considered overly familiar. Additionally, the argumentative essay topics may have felt less connected to the students' everyday experiences (refer back to Table 1), prompting more students to limit the use of reader engagement strategies.

Interestingly, compared to GSE A, GSE B featured a higher frequency of directives, which accounted for more than half of all such features in the main corpus (GSE). The vast majority of these were represented by the modal *should*, addressing the reader as the students' teacher, as exemplified in (19):

(19) The teacher's job should be this: give the opportunities to learner to expand their vocabulary. (GSE B essay)

The presence of directives – key dialogic features of argumentative writing (Jiang and Hyland 2025a) – also in 2 AGE B essays and 2 SSE B essays suggested the relevance of genre constraints on rhetorical choices.

The limited presence of appeals to shared knowledge, questions, and personal asides in GSE B, was in line with the findings of previous research on timed argumentative essay writing

(Zhao 2017). Nevertheless, personal asides and questions, though found in only 4 and 5 argument essays respectively, also appeared to set apart genuine student writing from LLM-generated essays. As exemplified in (20), asides were used to engage the reader (the teacher) by integrating a personal commentary into the text on the course topics discussed in class:

(20) the speed and efficiency of the use of ELF (and the rise of the globalized transnational culture that seems to come with it) is seen by most as a good trade off in the business community (GSE B essay)

Questions were used to arouse the reader’s curiosity and steer them towards the student’s personal interpretation of the title topic, as can be seen in (21):

(21) Today in all of the schools in the world English is taught (…). But is it correct how teachers explain English to the students ? This is actually a great question, that can look normal, but is not. (GSE B essay)

5.2.3 Discussion essays

Table 13 summarizes the stance and engagement features found in the three argument essays subcorpora.

Interactional items	**Frequency**			**Relative freq.**		
	AGE C	SSE C	GSE C	AGE C	SSE C	GSE C
Hedges	14	6	34	14941.30	6048.39	8145.66
Boosters	2	3	26	2134.47	3024.19	6229.04
Attitude	5	4	45	5336.18	4032.26	10781.03
Self-mentions	0	0	12	0.00	0.00	2874.94
STANCE TOT.	**21**	**13**	**117**	**22411.95**	**13104.84**	**28030.67**
Reader mentions	0	1	96	0.00	1008.06	22999.52
Directives	0	0	21	0.00	0.00	5031.15
Appeals to shared knowledge	1	0	1	1067.24	0.00	239.58
Asides	0	0	1	0.00	0.00	239.58
Questions	1	0	6	1067.24	0.00	1437.47
ENGAGEMENT TOT.	**2**	**1**	**125**	**2134.47**	**1008.06**	**29947.29**
	Document freq.			**Rel. Doc. freq.**		
	AGE C	SSE C	GSE C	AGE C	SSE C	GSE C
Hedges	4	4	11	100.00%	66.67%	61.11%
Boosters	2	2	12	50.00%	33.33%	66.67%
Attitude	3	3	17	75.00%	50.00%	94.44%
Self-mentions	0	0	11	0.00%	0.00%	61.11%
STANCE TOT.	**4**	**6**	**18**	**100.00%**	**100.00%**	**100.00%**
Reader mentions	0	1	17	0.00%	16.67%	94.44%
Directives	0	0	9	0.00%	0.00%	50.00%
Appeals to shared knowledge	1	0	1	25.00%	0.00%	5.56%
Asides	0	0	1	0.00%	0.00%	5.56%
Questions	1	0	5	25.00%	0.00%	27.78%
ENGAGEMENT TOT.	**2**	**1**	**17**	**50.00%**	**16.67%**	**94.44%**

Tab. 13: Interactional resources in the discussion essays

The overall pattern of stance and engagement features in discussion essays did not reveal any particularly striking differences in comparison to essays of the other types. However, AGE C featured relatively more hedges than both AGE A and AGE B. Considering that discussion essays are geared towards examining multiple sides of an issue and typically result in a more balanced perspective than argument essays, this characteristic of AGE C may be attributed to genre constraints. AGE C also featured a few attitude markers, which may likewise be linked to the nature of the discussion genre, allowing for greater expression of evaluative stance.

On the other hand, the relatively lower frequency of hedges in GSE C, compared to both AGE C and GSE B, may reflect the students' limited command of genre-specific conventions, particularly in relation to the discussion of essay topics (see Table 1). Given their close connection to course content and their role in stimulating active classroom discussion, these topics may have prompted students to prioritize their own perspectives, thereby overlooking the importance of addressing opposing viewpoints. Indeed, a close reading of the GSE C essays revealed a tendency among their authors to assert their own stance rather emphatically, with limited acknowledgment of alternative perspectives.

As previously noted, SSE C essays offered clear evidence of the hybrid nature of the SSE corpus as a whole. Some appeared to be re-elaborations of uncredited AI-generated output, while others merely resembled truncated sequences of arguments, possibly resulting from inattentive or time-pressured copying. The notably sparse use of stance features in SSE C – and particularly the complete absence of self-mentions, which mirrored the pattern observed in AGE C – further reinforced this interpretation.

A relatively lower frequency of self-mentions set apart GSE C from GSE A and GSE B, also possibly indicating the influence of genre (as in AGE C) and topic (as in GSE B), with GSE C essay titles addressing topics more detached from the students' everyday experience, although central to the course content. In GSE C, the reader mentions, directives, and questions, mirrored the overall engagement pattern of GSE. This confirmed the importance perceived by some students of involving the reader in the discussion as a conversational partner, as exemplified in (22) and (23):

(22) The question is, can foreign learners base their learning exclusively on standard English? We are going to see why varieties are actually so important, even nowadays. (...) To speak a language, we have to know the grammar but we also have to take into consideration the varieties of it (…) It's enough to think* about the differences between American and British English or to think about the various slangs used by the youth.

(* = *just think about*) (GSE C essay)

(23) Nowadays in our education system we study English language in its standard form. But is it useful in every situation? (GSE C essay)

On a final note, it must be observed that the only question featured in AGE C was in the heading that the LLM (DeepSeek-V4) had generated:

(24) The Global Spread of English: A Threat to Linguistic and Cultural Diversity ? (AGE C essay)

6. Discussion

The analysis set out in this paper suggests that interactive metadiscourse features may have limited explanatory power in distinguishing between AI-generated essays and essays genuinely written by students, even more so considering the size of the three corpora. That notwithstanding, on the one hand, the findings reported here are in line with previous research (Jiang and Hyland 2025b), revealing that LLMs' strengths are primarily evident in the use of interactive metadiscourse. In this investigation, LLMs were found to place strong emphasis on explicit cohesion, coherence and logical organization between ideas, demonstrating a preference for concise expression, particularly in compare-and-contrast essays. On the other hand, the data nevertheless highlighted that although intermediate EFL learners' writing falls short of an academic-like proficiency target, students are perhaps more sensitive than LLMs to the need to help readers interpret logical connections in arguments and guide them through the text.

Stance patterns did not prove particularly distinctive in separating genuine student writing from LLM-generated output. However, the findings nonetheless highlighted the more personal and individualized tone of the former. This was reflected in the higher relative frequency of attitude markers, boosters, and self-mentions in GSE compared to AGE. In contrast, the notably higher frequency of hedges in AGE indicated a more detached approach, with minimal engagement of the writer's persona.

Significantly, the findings indicated that LLMs showed clear limitations in audience modelling, as reflected in their constrained capacity to engage the reader. In contrast, genuine student writing, despite not yet meeting academic writing proficiency standards, was found to be more attuned than LLMs to the need to explicitly address and involve readers, as evidenced by the markedly higher frequency of engagement features in GSE, and particularly the consistent use of reader mentions. In sum, this research suggests that the absence of this dialogic quality is characteristic of essays produced by LLMs under zero-shot prompting, clearly distinguishing them from genuine EFL student writing.

It can be further suggested that the apparent addition of metadiscursive features by some students to LLM-generated text, in an attempt to evade detection, indicated that the characteristics identified in such essays were, in fact, perceived by students as lacking authenticity. It was noted that SSE exhibited patterns that alternately aligned with those of AGE and GSE, underscoring the hybrid nature of the suspect essays. It was suggested that the students' attempts to modify the LLM-generated texts to elude discovery, combined with partial or inattentive copying from such sources, likely produced texts that blended features of genuine and AI-generated writing. This interpretation was further supported using non-systematic, intuition-based content-related and linguistic criteria typically employed by teachers in detecting AI-generated text (De Wilde 2024). As noted earlier (5.1.3), SSE essays often contained content largely unrelated to course material, with some also exhibiting identical lexical patterns.

These findings broadly align with those from comparisons between LLM-generated text and English L1 student argumentative writing (Jiang and Hyland 2025a; 2025b; Mo and Crosthwaite 2025), emphasizing LLMs' limited capacity to reproduce human-like engagement features and pointing to the chiefly expository nature of LLM-generated text. However, it was also suggested that some students' understanding of academic conventions may have contributed to their reluctance to address the reader directly. Although the students had been told to ignore academic conventions advising against the use of personal pronouns and adverbs showing the writer's affective stance, it was previously suggested that not all of them appeared to have followed this guidance, with some deliberately minimizing use of engagement, possibly to maintain a more formal, non-conversational tone.

Furthermore, different interactional patterns observed across different essay types suggested possible constraints of genre on rhetorical strategies. Additionally, although this analysis did not consider the essay topic as a predictor variable, it was also suggested that topic choice may have influenced metadiscourse, with more abstract or impersonal topics prompting sparser use of stance and engagement features.

7. Conclusion

This study confirmed the author's initial intuition regarding the non-genuine nature of some essays submitted by EFL students. Specifically, with SSE alternating between patterns resembling AGE and GSE, the findings strongly suggested the hybrid nature of most SSE essays. Given the demonstrated unreliability of existing AI-detection tools, particularly with hybrid texts – and despite likely improvements over time – metadiscourse thus seems to constitute a viable framework through which the provenance of student essays can be assessed.

By extending comparative analyses of metadiscourse in LLM-generated text and human writing to an EFL context, this research suggested that, regardless of their target language proficiency, EFL learners' understanding of the writing task inherently involves consistent use of metadiscourse, and particularly interactional resources. Indeed, patterns of interactional features across the three corpora used in this investigation clearly indicated that acknowledging readers in the text and establishing rapport with them is a distinguishing characteristic of student writing that LLMs under zero-shot prompting apparently struggle to replicate.

The findings reported here align with those of previous research (Jiang and Hyland 2025a, 2015b; Mo and Crostwaithe 2025) carried out in English L1 contexts, suggesting that LLMs' training parameters tend to prioritize informational flow and adherence to academic formality conventions over interactional metadiscourse fostering reader engagement.

Raising EFL teachers' awareness of the characteristic features of LLM-generated text can support the detection of non-genuine student writing – although this task is often achievable through professional intuition and familiarity with students' academic profiles. Most importantly, such awareness can inform pedagogical practices aimed at fostering students' GenAI literacy. Learners stand to benefit from gaining a clearer understanding of how LLMs generate text, as well as the limitations inherent in their output. Accessibility of GenAI tools such as LLMs "does not necessarily translate into effective and meaningful use, especially for learning and instruction" (Cain 2024, 49), thus, the application of critical thinking remains essential for using these tools effectively and responsibly in academic settings.

7.1 Limitations and further research

Given its small-scale design, this research offered findings of limited statistical generalizability. Larger scale studies can yield more nuanced insights and more conclusive evidence. Also, this investigation analyzed students' essays written under time constraints and limited to approximately 200 words, which may have restricted the use of metadiscourse. Further studies might also look at metadiscourse features in longer, non-timed essays that have gone through revision and editing.

The AI-generated essays used for this analysis were created using zero-shot prompts. Further research may investigate the extent to which, through prompt engineering (Schulhoff et al. 2024), LLMs may be led to producing essays of various types that more closely resemble genuine student writing. One question that this study has indirectly raised is the extent to which LLMs can be fine-tuned to mimic human capabilities for audience evaluation and involvement.

Additionally, the focus of this research was on contrasting genuine student writing with LLM-generated texts, and not on benchmarking performance across multiple LLMs. However, only two LLMS were adopted, and possible differences in interactional features of written output among a wider and more varied range of LLMs were not considered. Further experimental investigations could usefully test the capabilities of different and possibly more specialized models.

Finally, given the rapid pace of advancement in GenAI tools, further research could also examine the capabilities of up-to-date AI-detection systems, assessing their reliability against corpus-assisted and qualitative analyses. An important line of inquiry would be to determine whether these approaches identify the same metadiscursive features as characteristic of genuine human writing or of LLM-generated output. Also, an open question remains as to how reliable such software can be when applied to hybrid texts.

Bionote

Marco Bagni is a research fellow at the Department of Communication and Economics of the University of Modena and Reggio Emilia. He has taught undergraduate courses in General English, English for Specific Purposes and English linguistics, and has worked for several years as a secondary school teacher in Italy. His research interests include the impact of GenAI in EFL pedagogy, Global English Language Teaching, English as a Lingua Franca, and English language variation. Among his recent publications, his monograph PhD dissertation: *Students Views and Attitudes Towards English and ELF in an Italian University* (Generis Publishing, 2024).

Works cited

Alexander, Katarzyna, Christine Savvidou and Chris Alexander. "Who Wrote this Essay? Detecting AI-Generated Writing in Second Language Education in Higher Education." *Teaching English with Technology* 23.2 (2023): 25-43.

Amirjalili, Forough, Masoud Neysani and Ahmadreza Nikbakht. "Exploring the Boundaries of Authorship: A Comparative Analysis of AI-Generated Text and Human Academic Writing in English Literature." *Frontiers in Education* 9 (2024): 9:1347421.

Barrot, Jessie S. "Using ChatGPT for Second Language Writing: Pitfalls and Potentials." *Assessing Writing* 57 (2023): 100745.

Basic, Zeljana, et al. "Better by You, Better than Me? ChatGPT-3 as Writing Assistance in Students' Essays." *arXiv* (2023). https://arxiv.org/abs/2302.04536. Last visited 20/06/2025.

Bishop, Lea. "A Computer Wrote this Paper: What ChatGPT Means for Education, Research, and Writing." *SSRN* (2023): 4338981. https://ssrn.com/abstract=4338981. Last visited 20/06/2025.

Borji, Ali. “A Categorical Archive of ChatGPT Failures.” *arXiv preprint* (2023): 2302.03494. https://arxiv.org/abs/2302.03494. Last visited 20/06/2025.

Cain, William. “Prompting Change: Exploring Prompt Engineering in Large Language Model AI and Its Potential to Transform Education.” *TechTrends* 68 (2024): 47-57.

Cogo, Alessia, Laura Patsko and Joanna Szoke. “Generative Artificial Intelligence and ELT.” *ELT Journal* 78.4 (2024): 373-377.

Creely, Edwin. “Exploring the Role of Generative AI in Enhancing Language Learning: Opportunities and Challenges.” *International Journal of Changes in Education* 1.3 (2024): 158-167.

De Wilde, Vanessa. “Can Novice Teachers Detect AI-generated Texts in EFL Writing?” *ELT Journal* 78.4 (2024): 414-422.

Frye, Brian L. “Should Using an AI Text Generator to Produce Academic Writing Be Plagiarism?” *Fordham Intellectual Property, Media & Entertainment Law Journal* 33.4 (2023): 946-968. https://ssrn.com/abstract=4292283. Last visited 20/06/2025.

Fyfe, Paul. “How to Cheat on Your Final Paper: Assigning AI for Student Writing.” *AI & Society* 38 (2023): 1395-1405.

Herbold, Steffen, et al. “A Large-Scale Comparison of Human-Written Versus ChatGPT-Generated Essays.” *Scientific Reports* 13.1 (2023): 18617.

Ho, Victor and Cissy Li. “The Use of Metadiscourse and Persuasion: An Analysis of First Year University Students’ Timed Argumentative Essays.” *Journal of English for Academic Purposes* 33 (2018): 53-68.

Huang, Jerry. “Engineering ChatGPT Prompts for EFL Writing Classes.” *International Journal of TESOL Studies* 5.4 (2023): 73-79.

Hyland, Ken. *Metadiscourse: Exploring Interaction in Writing*. London: Continuum, 2005.

---. “Stance and Engagement: A Model of Interaction in Academic Discourse.” *Discourse Studies* 7.2 (2005): 173-192.

Jiang, Feng Kevin and Ken Hyland. “Does ChatGPT Write Like a Student? Engagement Markers in Argumentative Essays.” *Written Communication* 42.3 (2025a): 463-492.

---. “Rhetorical distinctions: Comparing Metadiscourse in Essays by ChatGPT and Students.” *English for Specific Purposes* 79 (2025b): 17-29.

Kohnke, Lucas, Benjamin Luke Moorhouse and Di Zou. “ChatGPT for Language Teaching and Learning.” *RELC Journal* 54.2 (2023): 537-550.

Mo, Zishan and Peter Crosthwaite. “Exploring the Affordances of Generative AI Large Language Models for Stance and Engagement in Academic Writing.” *Journal of English for Academic Purposes* 75 (2025): 101499.

Moorhouse, Benjamin Luke, Marie Alina Yeo and Yuwei Wan. "Generative AI Tools and Assessment: Guidelines of the World's Top-Ranking Universities." *Computers & Education Open* 5 (2023): 1-10.

Ramazani, Alireza, Houman Bijani and Mohammad Reza Oroji. "Comparative Analysis of AI vs. Human Feedback Effects on IELTS Candidates' Writing Performance." *Journal of Foreign Language Teaching and Translation Studies* 10.1 (2025): 17-40.

Rong, Hui and Charlene Chun. "Digital Education Council Global AI Student Survey 2024." *Digital Education Council* (2024). https://www.digitaleducationcouncil.com/post/digital-education-council-global-ai-student-survey-2024. Last visited 20/06/2025.

Schulhoff, Sander, et al. "The Prompt Report: A Systematic Survey of Prompting Techniques." *arXiv preprint* (2024): 2406.06608. https://arxiv.org/abs/2406.06608. Last visited 20/06/2025.

Su, Yanfang, Yun Lin and Chun Lai. "Collaborating with ChatGPT in Argumentative Writing Classrooms." *Assessing Writing* 57 (2023): 100752.

Teng, Mark Feng. "Metacognitive Awareness and EFL Learners' Perceptions and Experiences in Utilising ChatGPT for Writing Feedback." *European Journal of Education* 60.1 (2025): e12811.

Warschauer, Mark, et al. "The Affordances and Contradictions of AI-Generated Text for Writers of English as a Second or Foreign Language." *Journal of Second Language Writing* 62 (2023): 101071.

Woo, David J., et al. "Exploring AI-Generated Text in Student Writing: How Does AI Help?" *Language Learning & Technology* 28.2 (2024): 183-208.

Yoo-Jean, Lee. "Can My Writing Be Polished Further? When ChatGPT Meets Human Touch." *ELT Journal* 78.4 (2024): 401-413.

Yoon, Hyung-Jo. "Interactions in EFL Argumentative Writing: Effects of Topic, L1 Background, and L2 Proficiency on Interactional Metadiscourse." *Reading and Writing* 34.3 (2021): 705-725.

Zhao, Cecilia Guanfang. "Voice in Timed L2 Argumentative Essay Writing." *Assessing Writing* 31 (2017): 73-83.