IJE
International Journal of Education

Journal homepage: https://ejournal.upi.edu/index.php/ije/index

A COMPARATIVE ANALYSIS OF AI AND HUMAN EVALUATION OF HEDGES AND BOOSTERS IN STUDENT ACADEMIC WRITING

Farida Hidayati
English Language and Literature Study Program, Universitas Pendidikan Indonesia, Bandung, Indonesia
***Corresponding author's E-mail address: faridahidayati@upi.edu**

ABSTRACT

Hedges (such as, might, perhaps) and boosters (such as, clearly, undoubtedly) are central epistemic devices and epistemic pragmatics devices in academic writing. However, student writers often find it tough to engage in their certain participation. Hence, it is significant to evaluate how such AI models such as ChatGPT-4.5 compare with human teachers on assessing these epistemic features. This study aims to compare the scores and evaluative feedback provided by an AI model (ChatGPT-4.5) and a human writing instructor in assessing the deployment of hedges and boosters in four undergraduate argumentative essays. The mixed methods were used to analyze the feedback from ChatGPT and an experienced writing instructor and their ratings of each essay across 10 components based on Hyland's framework on a 5-point scale. The Mann-Whitney U test showed that there was no statistically significant difference in overall scores between the AI and human rater (U-statistic: 59.5000, P-value: 0.4644), indicating general alignment. However, differences were shown at the component level: AI was less variable when it came to identifying hedges and boosters, whereas the instructor added greater context to his comments regarding the appropriateness of use. Both raters were found to correlate moderately for stance evaluation. From a qualitative perspective, thematic analysis revealed an AI's generic phrase usage within a limited context and a teacher pedagogy-grounded, rhetorically informed comment. It turns out that students also stated a more constructive approach for the teacher due to clarity and helpful relevance.

ARTICLE INFO

Article History:
Received 9 May 2025
Revised 14 Jun 2025
Accepted 29 Jun 2025
Available online 30 Jun 2025

Keywords:
Academic writing; AI evaluation; boosters; evaluation; hedges

To cite this paper (in APA style):
Hidayati, F. (2025). A comparative analysis of AI and human evaluation of hedges and boosters in student academic writing. *International Journal of Education, 18*(2), 155-166. https://doi.org/10.17509/ije.v18i2.85627

1. INTRODUCTION

Hedges and boosters are important in academic writing as they enable writers to express varying degrees of certainty, doubt, and emphasis in their argumentation. Not only do these meta discourse markers establish the writer's epistemic stance, but they also function interpersonally to orient the text to the conventions of academic discourse communities (Shen, 2023). Studies have shown that EFL learners, and especially in the context of ASEAN, tend to overuse boosters at the expense of hedges, which mirrors cultural communication styles and results in text that is read as overbearing or not nuanced enough (Ningrum et al., 2024). The same difficulties have been detected among L2 writers elsewhere in the world, for instance, Yemeni scholars who used hedges and boosters in their academic writing very infrequently (Al-mudhaffari et al., 2020). These inclinations indicate the greater cross-cultural variation in the utilization of hedging and boosting devices (Dontcheva-Navratilova, 2016). Effective use of these devices not only enables the transmission of ideational information but also enables the writer to have an audience-aware and persuasive voice in engagement with the academic world. The pedagogy of academic writing therefore must focus on strategic deployment of hedges and boosters, with corresponding feedback to facilitate students in moving towards the development of an audience-aware and balanced academic voice.

The introduction of AI writing tools such as ChatGPT has offered new possibilities in teaching writing and assessing writing, particularly in English as a Foreign Language (EFL) context. The tool can offer real-time feedback and assess multiple dimensions of the linguistic features of students' writing, which offers momentum to the quality of writing improvement (Xiao et al., 2025). Studies have shown that AI-assisted tools can help enhance students' motivation, confidence, and awareness of academic discourse, particularly through iterative feedback that supports self-revision (Song & Song, 2023; Bok & Cho, 2023). Nevertheless, recent studies indicate that AI falls behind in recognizing advanced and subtle characteristics of university language, for instance, epistemic stance markers such as boosters and hedges, which are highly context-dependent and insipid in nature (Algaraady & Mahyoob, 2023; Steiss et al., 2024). Although AI can complement traditional instruction, it is less capable than human teachers in giving piercing, pointed comments and highlighting the most important points of academic writing (Algaraady & Mahyoob, 2023; Steiss et al., 2024). Research also notes that while AI-generated feedback provides linguistic accuracy, human feedback often contributes to deeper rhetorical awareness and critical thinking (Gayed et al., 2022; Kim et al., 2023). Specifically, there is no significant difference in learning gains between feedback from AI and human intelligences for EFL students, which promises feasibility for application (Escalante et al., 2023). Despite the potential of AI, its appropriacy and quality of feedback are still a concern, particularly for EFL students who struggle with hedges and boosters. Hence, scholars call for a blended pedagogy that draws upon the merits of AI efficiency and human tact in feedback provision, as well as further research to tailor AI feedback to EFL learners' discourse needs and to address ethical considerations in AI-assisted learning (Bok & Cho, 2023; Nguyen Thi Thu, 2023; Song & Song, 2023; Xiao & Zhi, 2023; Xiao et al., 2025).

In line with this, research on formative feedback highlights not only the content but also the linguistic delivery as equally essential for enhancing learner engagement and learning outcomes. Shute (2008) characterizes effective formative feedback as non-evaluative, supportive, timely, and specific in order that learners may alter behavior and improve performance. Other than the information content itself, linguistic form through which feedback is delivered, most importantly through pragmatic devices like hedging and boosting, has powerful impacts on reception and uptake. For example, Ryoo (2023) provides that hedged sentence feedback like "perhaps consider" or "it would be better if…" can trigger L2 writers' revision while maintaining learners' confidence. However, Ryoo also cautions that too implicit and generic feedback can turn out to be a barrier to revision accomplishment and lead to declining student confidence. Wingate (2010) established that students who responded to rich feedback improved in previously targeted areas, and that students who unattended or skewed feedback repeated the same mistakes. Wingate additionally stressed that student uptake of feedback depends on students' motivation, engagement with their subject of study, and their writer identity. The findings highlight the importance of pragmatically aware and culturally sensitive feedback approaches, especially in EFL environments, where linguistic as much as cultural expectations could be other than those that are taken for granted by the providers of feedback.

Previous studies have shown that EFL writers tend to fare poorly with using hedges and boosters appropriately. They tend to overuse boosters such as "always" or "completely" but underuse hedges such as "might" or "possibly," resulting in writing that is not subtle or comes across as quite forceful (Ningrum et al., 2024; Al-mudhaffari et al., 2020). These advancements are typically accounted for in terms of educational and cultural reasons, as well as a lack of overt instruction of epistemic stance marking. In parallel, uses of AI programs such as *ChatGPT* to aid with writing instruction have been widely studied. Some research shows that feedback from AI can be rich and effective (Solak, 2024; Tran et al., 2023), but it lacks contextual sensitivity and emotional nuance. Human raters and AI have been reported in certain studies to show moderate correlation when marking aspects of writing (Geçkin et al., 2023), but when grading higher-order skills and rhetorical such as stance, human raters are more accurate and precise in scoring (Steiss et al., 2024; Karademir Coşkun, 2024).

Aside from the growing body of studies on AI-aided writing assessment and hedging and boosting use by EFL writers, very little research has compared directly how AI and human examiners assess such specific linguistic markers. Most prior research tested general writing quality or more general areas of feedback without closely examining whether AI would be able to effectively detect and assess epistemic stance markers such as hedges and boosters. While others have compared essays produced by AI and human essays in terms of structural or grammatical accuracy (Charpentier-Jiménez, 2024), not many have addressed pragmatic features

DOI: https://doi.org/10.17509/ije.v18i2.85627
p- ISSN 1978-1342 e- ISSN 2442-4730

such as stance. Geçkin et al. (2023) found a moderate and statistically significant correlation between *ChatGPT-3.5* and a human rater in evaluating L2 college writing, but the analysis did not account for individual characteristics such as hedging and boosting. Similarly, Steiss et al. (2024) concluded that while *ChatGPT* can match or even surpass human grading in the instance of standards-based grading, the opposite is true when it comes to grading rhetorical nuance. The study cites a primary limitation with particular reference to formative writing assessment in the university context where rhetorical devices are most valuable in argument and in an author's position.

To fill this research gap, the present study aims to compare an AI model (*ChatGPT-4.5*) and a human writing instructor's evaluative score and comments regarding the application of hedges and boosters in undergraduate EFL argumentative essays. Specifically, the study investigates whether AI assessment equals human assessment in quantitative ratings and qualitative comments, with regards to the rhetorical function served by hedging and boosting. The primary objective of this study is to determine the degree of comparability between *ChatGPT-4.5*'s assessments and human raters' assessments regarding the identification and analysis of hedges and boosters in EFL student essays. By narrowing down to these epistemic stance markers, the research attempts a more precise contribution to the knowledge of how AI can be implemented in writing evaluation. The findings will guide pedagogy in writing, academic writing curriculum design, and ethical application of AI in formative feedback systems. Findings can also guide developers of AI on how to tune channels of feedback for more effective rhetorical and pragmatic evaluation in EFL contexts.

While ethical and pedagogical issues remain important in broader AI-in-education discussions, the current research is not interested in those aspects. It is interested in a relative comparison of hedges and boosters by AI and human assessors in EFL academic discourse.

2. METHOD

2.1. Research Design

This study adopted a convergent mixed methods design, where qualitative and quantitative methods were fused to provide a comprehensive evaluation of how hedges and boosters are assessed in the academic writing of students by both an AI system (*ChatGPT-4.5*) and a human rater (a writing teacher). The quantitative strand involved marking the use of hedges and boosters among the students through 10 elements supported by Hyland's (2005) model

Table 1. Evaluation Components adapted from Hyland (1998)

No.	Evaluation Components	Definitions
1.	Accuracy of Identification	Correctly recognizing all instances of hedges, boosters, and related epistemic devices.
2.	Appropriateness of Usage	Use of hedges and boosters that fits the academic context and maintains communicative clarity.
3.	Effectiveness of Stance	The extent to which hedges and boosters express the writer's intended level of certainty, caution, or emphasis.
4.	Variability and Range of Devices	Diversity in the types of hedges and boosters used, avoiding repetition and showing command of pragmatic options.
5.	Interactional Function	How well the writer engages with the reader, using hedges and boosters to negotiate stance, create solidarity, or manage face.
6.	Rhetorical Impact	The contribution of hedges and boosters to the overall persuasiveness, coherence, and flow of the argument.
7.	Contextual Sensitivity	Sensitivity to disciplinary norms and audience expectations in the use of epistemic devices.
8.	Grammatical and Lexical Accuracy	Correct grammatical form and lexical choice in the use of hedges and boosters.
9.	Balance Between Certainty and Caution	Ability to appropriately balance boosting confidence with hedging uncertainty to strengthen the argument

DOI: https://doi.org/10.17509/ije.v18i2.85627
p- ISSN 1978-1342 e- ISSN 2442-4730

		without overstating.
10.	Pedagogical Feedback Quality	Clarity, relevance, and constructiveness of comments and suggestions related to hedges and boosters in feedback.

Each aspect was graded independently on a numeric scale, enabling close comparison between AI and human marking. The study design allowed for triangulation of results and enabled deeper understanding of evaluative patterns and interpretive inclinations. By analyzing the convergence between AI and human marking, the study seeks to determine the viability of AI as a method of assessing pragmatic features of academic writing.

Table 2. Rubric of the Use of Hedges and Boosters

The following rubric, which lists the essential elements used to evaluate how well authors use hedging and boosting devices in academic writing, is explained in detail below:

Number	Component	Description	Scoring Guide (1–5)
1	Accuracy of Identification	Correctly spotting hedges and boosters	1 = Mostly missed; 5 = All correctly identified
2	Appropriateness of Usage	Contextually suitable and meaningful use	1 = Inappropriate; 5 = Fully appropriate
3	Effectiveness of Stance	Expresses intended epistemic stance clearly	1 = Ineffective; 5 = Very effective
4	Variability and Range	Use of diverse hedges and boosters	1 = Very limited; 5 = Wide and varied
5	Interactional Function	Engages reader and manages interpersonal aspects	1 = Poor; 5 = Excellent
6	Rhetorical Impact	Enhances persuasiveness and coherence	1 = Weak; 5 = Strong
7	Contextual Sensitivity	Aligns with disciplinary and audience expectations	1 = Misaligned; 5 = Fully aligned
8	Grammatical and Lexical Accuracy	Correct grammar and word choice in hedges and boosters	1 = Many errors; 5 = No errors
9	Balance Between Certainty and Caution	Uses hedges and boosters to balance confidence and caution	1 = Poor balance; 5 = Excellent balance
10	Pedagogical Feedback Quality	Clarity and usefulness of evaluative comments	1 = Unclear/unhelpful; 5 = Clear/constructive

The table above presents a comprehensive rubric for assessing the application of hedges and boosters across ten distinct components. Each component is evaluated on a five-point scale, with a score of 1 indicating inadequate performance and a score of 5 reflecting highly effective and sophisticated use. What follows is a detailed explanation of each criterion. By facilitating the strategic deployment of hedging and boosting devices, this rubric functions as both an evaluative instrument and a pedagogical framework, enabling writers and educators to critically examine and enhance the rhetorical efficacy of academic texts.

2.2. Participants

Research participants were composed of three groups: four student writers, one human rater, and a machine model.

Student Writers: Four undergraduate students (two males, two females) aged 19–21 were chosen purposively from the English Language and Literature Study Program of a public university in Bandung, Indonesia. They had all completed at least one academic writing course and were familiar with the conventions of formal, argumentative writing. Purposive sampling was used to ensure that the students had prior knowledge of epistemic stance devices such as hedges and boosters so that their writing could be more informatively analyzed. Gender and proficiency difference were taken into consideration to see the possible effect of different

DOI: https://doi.org/10.17509/ije.v18i2.85627
p- ISSN 1978-1342 e- ISSN 2442-4730

student backgrounds on the use of these pragmatic features. While the small number of student writers' limits generalizability, it is consistent with the exploratory, in-depth nature of this qualitative-comparative case study. Voluntary participation was witnessed, and informed consent was provided by all the students.

Human Evaluator: The human rater was a male senior lecturer in academic writing, aged 54, with over 20 years of teaching experience and a Ph.D. in applied linguistics. He was selected due to his extensive teaching experience in rhetorical analysis, expression of stance, and formative assessment of writing. The rater was requested to rate and comment on the essays on a typical rubric for three major categories of hedges and boosters: (1) identification accuracy, (2) usage appropriacy, and (3) stance effectiveness.

AI Evaluator: The AI model utilized in this study was ChatGPT-4.5, accessed through the official OpenAI ChatGPT website (2025 version). It was selected because of its state-of-the-art natural language processing capabilities, including rhetorical analysis and feedback generation. The AI was walked through carefully designed prompts to perform evaluations that were identical in scope and focus to those performed by the human scorer. All outputs were saved for later analysis.

2.3. Data Collection Procedure

To enable comparison that is systematic and equitable, data collection entailed several steps. The primary data consisted of academic articles written by the student participants as part of their final-year graduation requirements in the English Language and Literature Study Program. Each student had written an article of 5,000–7,000 words intended for submission to nationally accredited journals (at least SINTA level). These texts represented authentic academic writing tasks and provided a suitable corpus for analyzing the use of hedges and boosters. The anonymization of all the student essays was the first step to remove names and any recognizable details, thus maintaining the evaluation unbiased for human and AI evaluators alike. There were two evaluators: The Human Evaluator and The AI Evaluator.

The Human Evaluator was given the anonymized essays along with a complete evaluation rubric. The rubric included criteria that are based on three primary epistemic stance marker use areas:

- Identification (whether the hedges and boosters were well-placed)
- Appropriateness (to what extent the markers suited the context and content)
- Effectiveness (to what extent the stance was conveyed to the reader)

The teacher marked each essay by underlining useful examples, writing brief marginalia, and awarding scores (on a scale of 1–5) against each criterion. The AI Evaluator, ChatGPT-4.5 (accessed via OpenAI's 2025 ChatGPT platform), was given the same texts along with standardized, structured prompts. These prompts mirrored the rubric instructions and were issued sequentially to avoid overloading or ambiguity. For each essay, the prompt included:

> *"Identify all the hedges and boosters in the text below. Score each on a 1–5 scale for (1) accuracy of identification, (2) appropriateness of use, and (3) effectiveness of stance. Explain your scores. This controlled prompting procedure ensured that human and AI evaluators worked under comparable evaluation conditions."*

After evaluation, qualitative remarks and all the scoring forms by both the raters were collected for thematic analysis and quantitative analysis. Moreover, in the hope of obtaining the student voice, short semi-structured interviews (15–20 minutes each per respondent) were conducted with each of the four student authors. The interviews asked them about their views concerning the feedback that they had received, by both the AI and teacher, on their clarity, usefulness, and general preference. All the interviews were tape-recorded with consent, transcribed verbatim, and analyzed thematically.

2.5. Data Analysis

The study used both quantitative and qualitative approach to analyze the data gathered from AI-and human-provided ratings, and from interviews with the students. The quantitative investigation focused on three aspects of evaluation most crucial for epistemic stance: (1) precision in hedge and booster identification, (2) appropriateness of use, and (3) rhetorical effectiveness in conveying stance. Both the human instructor and ChatGPT-4.5 were instructed to rate these components on a 1–5 scale for each student essay. As the resulting scores were ordinal and did not follow a normal distribution, the Mann-Whitney U test was selected as an appropriate non-parametric statistical method to compare the distributions of scores assigned by each evaluator. Statistical significance was determined at the $p < 0.05$ level. All computations were conducted with the Python programming language, i.e., by using scipy.stats for statistical testing and pandas for data manipulation. A convenient graphical user interface was implemented using the tkinter library for visualization and exploration of

DOI: https://doi.org/10.17509/ije.v18i2.85627
p- ISSN 1978-1342 e- ISSN 2442-4730

the score data. All the development and analysis were carried out in the Visual Studio Code (VS Code) environment to ensure reproducibility and transparency.

To complement the quantitative results, qualitative content analysis of human rater and ChatGPT's feedback comments were conducted. The use of open coding identified each of the comments on dimensions of pedagogical depth, specificity, contextual salience, and tone. Initially, separate segments of feedback were coded and progressively merged into more general themes such as directive versus suggestive phrasing, superficial versus deep content feedback, and generic versus context-sensitive engagement.This thematic analysis enabled a close examination of the differences in evaluative discourse styles between the AI model and the human teacher, especially regarding how each conveyed pragmatic and interpersonal dimensions of feedback.

To further triangulate the findings and understand the reception of feedback, each of the four student participants was invited to a semi-structured interview lasting approximately 15–20 minutes. The interviews questioned students about their response to the clarity, usefulness, and emotional impact of feedback from the two raters. Sample guiding questions included: "Which feedback did you find more useful to revise?", "Did you find the human or AI feedback more encouraging or informative?", and "How confident do you feel to use hedges and boosters after getting the feedback?" Interview responses were audio-recorded, transcribed, and coded thematically. Themes were generated according to perceived usefulness, motivational effect, and clarity, working to place the students' feedback requirements in the context of their learning requirements and levels of engagement.

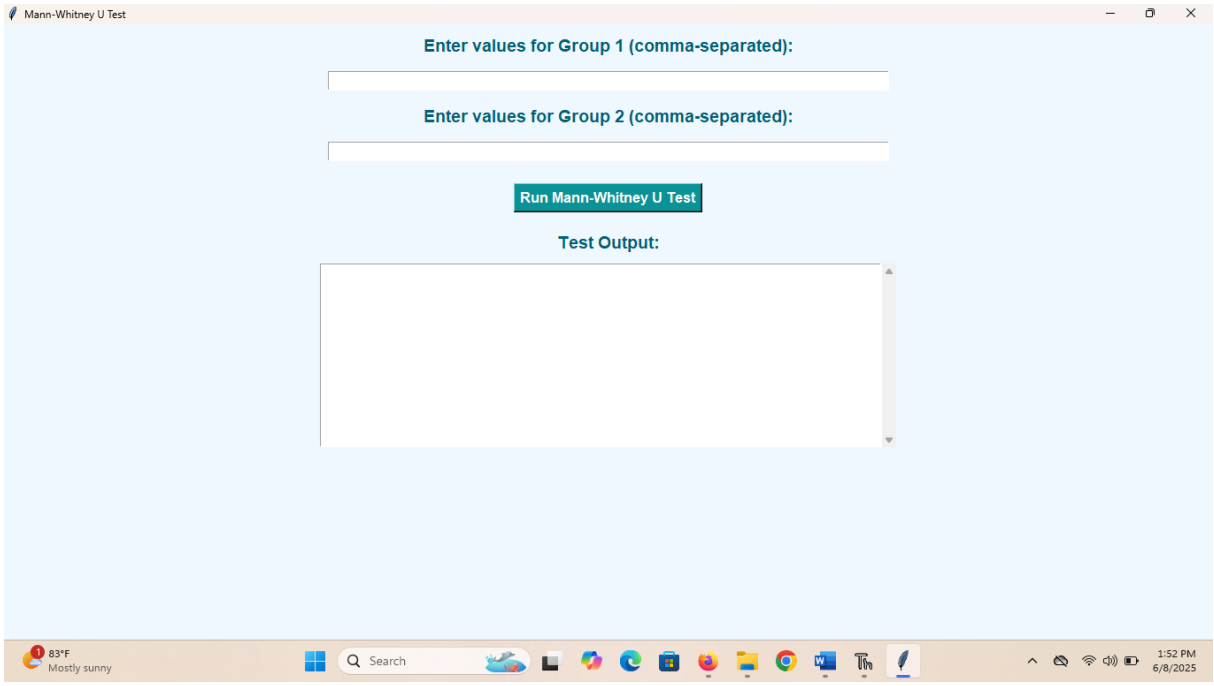

Figure 1. The Mann-Whitney U Test Machine

For Qualitative Analysis, all teacher markings (and AI output) were theme coded. Open coding was used to look for patterns in comments, maintaining depth of explanation, pedagogical clarity, reference to context, and tone of comments as focus areas. Codes were developed iteratively into themes showing variations in AI vs. human style of comments.

Additionally, attitudes from students were captured in brief follow-up semi-structured interviews (15–20 minutes each student) asking for opinions on the clarity, usefulness, and preference of teacher versus AI feedback. The interviews took place, were transcribed, and coded thematically.

3. RESULTS AND DISCUSSION

3.1. Quantitative Comparability of ChatGPT-4.5 and Human Evaluator in Scoring Hedges and Boosters

This chapter answers Research Question 1 (RQ1): To what extent are the performances of a human teacher and ChatGPT-4.5 similar in assessing hedges and boosters in EFL undergraduate essays? Table 3 presents the mean scores of the two raters in ten components. Overall, there is a high degree of consistency, with AI and the human rater assigning similar scores in most of the categories. For example, they both scored Accuracy of Identification, Effectiveness of Stance, and Grammatical and Lexical Accuracy the same (mean score = 4.0–4.5). Minor differences are registered in Appropriateness of Usage and Interactional Function, where the human rater gave higher scores.

DOI: https://doi.org/10.17509/ije.v18i2.85627
p- ISSN 1978-1342 e- ISSN 2442-4730

Overall Score Comparisons

The mean scores assigned by both evaluators across all components and essays are summarized in Table 1.

Table 3. Mean Scores by Evaluator and Component (n = 4 essays)

No.	Component	Human Evaluator	ChatGPT-4.5
1	Accuracy of Identification	4.00	4.00
2	Appropriateness of Usage	4.25	4.00
3	Effectiveness of Stance	4.00	4.00
4	Variability and Range	3.75	3.75
5	Interactional Function	4.50	4.25
6	Rhetorical Impact	4.00	4.00
7	Contextual Sensitivity	4.25	4.00
8	Grammatical and Lexical Accuracy	4.50	4.50
9	Balance Between Certainty and Caution	4.00	4.00
10	Pedagogical Feedback Quality	4.75	4.50

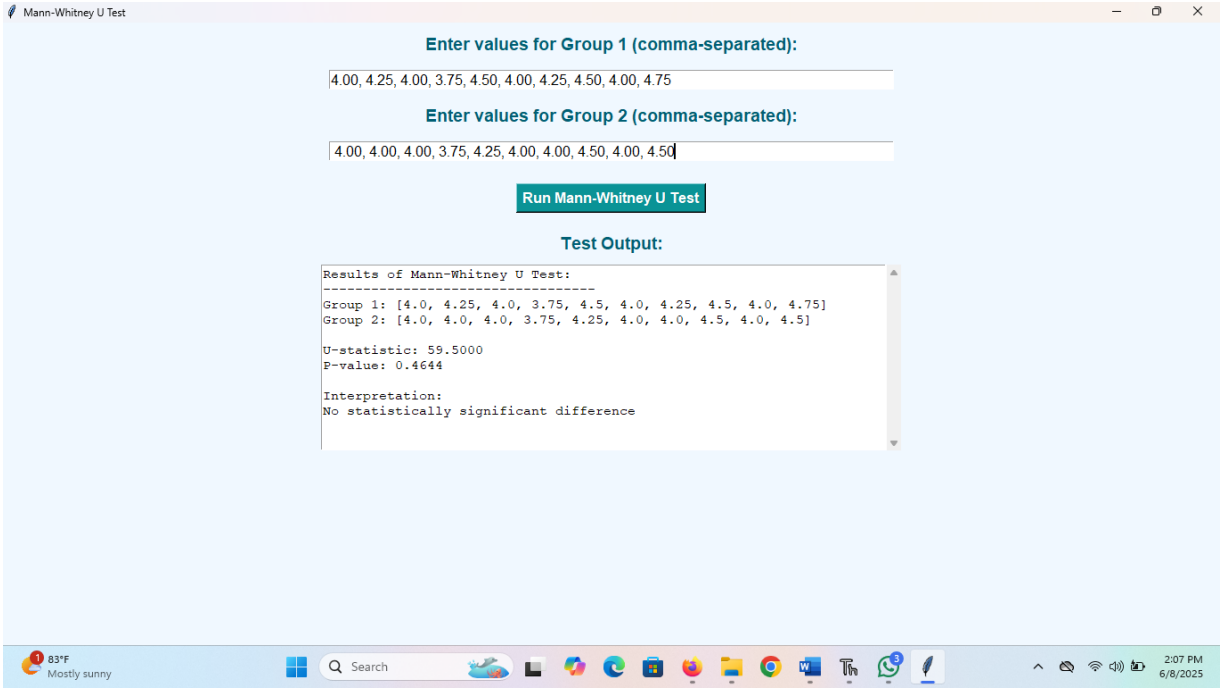

Figure 2. A Display of the results of the Statistical Computation Using the Mann-Whitney U Test

A Mann-Whitney U test was conducted to statistically examine the differences between the two raters. The results indicated no statistically significant difference (U = 59.5000, p = 0.4644), supporting the conclusion that AI and human evaluations were generally consistent. Despite this, subtle tendencies emerged:

- ChatGPT-4.5 outperformed in raw identification, especially with ambiguous modals such as may and seems.
- The human rater gave slightly higher scores for rhetorical appropriateness, showing more tolerance for vague or strategic hedges.

Conclusion to RQ1: ChatGPT-4.5 offers a reliable, though not identical, alternative to human scoring in hedge/booster assessment. It is especially competent at detection but may lack the rhetorical nuance of human raters when assigning appropriateness judgments.

3.2. Qualitative Findings: Qualitative Differences in Feedback Characteristics

DOI: https://doi.org/10.17509/ije.v18i2.85627
p- ISSN 1978-1342 e- ISSN 2442-4730

3.2.1 Thematic Analysis of Evaluator Feedback

This section addresses Research Question 2 (RQ2): What are the differences in the style, tone, and pedagogical richness of responses by ChatGPT-4.5 versus a human teacher? Thematic analysis revealed four broad differences between the feedback styles:

Theme 1: Depth of Interpretation

The human teacher consistently wove explanation and rhetorical purpose together in his comments, while ChatGPT provided more shallow comments (e.g., "The use of might shows uncertainty") without contextual explanation.

> *It looks like code-switching might have several pragmatical purposes in bilingual classrooms, particularly where cultural identity often shifts. Certain scholars argue that it merely reflects linguistic deficiency; this is a belief that it could reflect a strategic way to create solidarity or give meaning. This phenomenon might be subject to the social roles of interlocutors and the particular interactional context. So, probably that code-switching is more than only a communicative bridge; it likely supports cultural belonging and discourse coherence as well.* **(Student 1-Male)**

The excerpt demonstrates the student's ability to integrate linguistic observation with sociocultural interpretation, moving beyond descriptive commentary. By linking code-switching to identity construction and interactional context, the student shows an awareness of its pragmatic and symbolic functions rather than treating it as mere linguistic deficiency. This reflects a higher level of rhetorical purpose and critical engagement compared to ChatGPT's surface-level explanation.

Theme 2: Pedagogical Sensitivity

Human feedback often included actionable suggestions tailored to drafts. For instance, students were instructed to "make this claim stronger to respond more in the assertive tone of the paragraph." ChatGPT was more strictly bound to conventional grammar and structure conventions.

> *Mary Shelley's employment of natural imagery in Frankenstein arguably amplifies the moral ambiguity of her characters. For example, the persistent presence of the sublime landscape speaks to a sort of psychological space between awe and terror. While some critics insist that Victor Frankenstein is entirely responsible for the creature's ultimate downfall, it is probably more nuanced to see Frankenstein as a character trying to comprehend the limits of human knowledge and ambition. In this regard, readers might be willing to reassess his behaviour, not as reckless arrogance, but rather as a complicated reaction to the ideals of the Enlightenment and his own guilt. Ultimately, Shelley complicates moral assessments so that it is difficult to pinpoint the culpability.* **(Student 2-Female)**

The excerpt shows how the student incorporated human feedback to strengthen argumentation and adopt a more assertive academic tone. The student moves from simple literary description to analytical interpretation, situating Victor Frankenstein's actions within broader philosophical and moral frameworks. This indicates an increased rhetorical awareness and critical depth that aligns with the targeted feedback on developing stronger argumentative voice.

Theme 3: Lexical Breadth

Even though ChatGPT marked a wider variety of hedging expressions, it would sometimes misidentify neutral sentences as hedges/boosters in isolation. The human rater was more conservative but correct on pragmatic marking.

> *The social and linguistic variations in urban youth speech reflect changing group identities among urban youth. Non-standard forms are sometimes adopted not purely for the sake of opposing mainstream value systems but also to signal acceptance in an in-group membership. The linguistic items could, therefore, be influenced by media exposure, peer networks, and local cultural capital. By the same token, apparently neutral phrases like "you know" or "kind of" may adopt subtle social meanings. Such markers may not only mitigate the statements but actually reinforce solidarity or distance, depending on whether they are framed in the interaction as collaborative or confrontational..* **(Student-3- Female)**

By emphasizing the relational functions of expressions like "you know" or "kind of," the student demonstrates awareness of how such markers construct social meaning within interaction. This depth contrasts with ChatGPT's tendency to classify hedges mechanically, showing the human rater's greater sensitivity to discourse pragmatics.

Theme 4: Pragmatic Function and Contextual Awareness

The teacher tended to equate the use of hedge/booster with broader argumentative strategy, which the AI rarely did. This is consistent with previous research (e.g., Yoon et al., 2023), where AI output is less mindful at the discourse level.

DOI: https://doi.org/10.17509/ije.v18i2.85627
p- ISSN 1978-1342 e- ISSN 2442-4730

This use of pragmatic markers in the essay is indicative of the writer's effort to negotiate authority and politeness in an academic context. Phrases such as "it might be argued" or "this could suggest" not only reduce certainty but also engage reader perspective and acknowledge alternative views. This strategic language choice could also balance claims of confidence with claims of humility, which is of particular relevance in an academic context where knowledge claims are liable to challenge. Therefore, it is plausible that the writer uses these markers to hedge epistemic commitment while attempting to engage in a cooperative dialogue with the audience.

> *One could argue that code-switching is an important part of bilingual classrooms, beyond just being a language shortcut. Some researchers argue that code-switching is just an indication of language deficiency, but it might be more accurate to think about code-switching as a way that speakers can establish group identity and solidarity. Certainly, code-switching does have distinct functions reflecting a range of social contexts and interlocutors. In many classroom interactions, it clearly supports both cultural belonging and pragmatic coherence, allowing speakers to participate in multiple linguistic and cultural norms in real time. This highlights that code-switching is not only a communicative bridge, but also an involved practice within multiple and interwoven social layers.* **(Student 4-Male)**

While both the evaluators provided grammatically sound feedback, the human instructor's comments were more rhetorically rich, pedagogically sensitive, and contextually aware qualities that are most critical in formative feedback for advanced EFL writing.

3.2.2 Student Perceptions of Feedback

All four participants were interviewed after reading the feedback given by the human instructor and AI (ChatGPT-4.5). Through thematic analysis of their responses, some significant themes emerged:

Clarity and Usefulness: Students across the board noted that teacher feedback was clearer and more helpful, particularly as it revealed why a particular hedge or booster performed or failed in the rhetorical context. A student stated, for example, "The teacher's feedback not only told me what was wrong, but why I had to do it differently to improve my argument."

AI Helpful: Two students appreciated the speed and the completeness of the AI feedback, observing that it quickly identified many hedges and boosters. They did also find the AI comments a bit "*too repetitive*" and "*too general*," though, and lacking the subtle detail that a human tutor could provide. A volunteer explained, "*ChatGPT picked out many hedges, but sometimes it felt like it was repeating the same kind of phrase without really entering my true writing style."*

Source of Feedback Preferred: All the students favored human feedback for their final papers, valuing the depth and contextualization it offered. They were, however, happy to use ChatGPT as a secondary means for proofreading early drafts, especially when working individually or in looking for immediate feedback. According to one student, "*I like using ChatGPT when I'm stuck or need ideas quickly, but for the final paper, I pay more attention to what the teacher says*."

3.2.3 Interview Excerpts and Analysis

Excerpt 1: Clarity and Usefulness

Interviewer: *"How did you apply teacher feedback to improve your hedges and boosters?"*

Student 1: *"The instructor didn't just tell us 'this is a hedge' or 'this doesn't work.' They told us why it works there, such as showing how a hedge can soften a claim when evidence isn't sufficient. That helped me see the bigger picture a lot."*

Analysis: The response exhibits how human feedback indicates pedagogical value through the connection of linguistic features to rhetorical functions, which students appreciated for enriched learning.

Excerpt 2: AI Usefulness and Limitations

Interviewer: *"Was your experience with the AI comments compared to that of your teacher?"*

Student 2: *"ChatGPT was rapid and caught many hedges and boosters, which was useful for the quick checking. But sometimes the comments sounded generic — like it didn't understand my point. It just used to pick out the words and not define their purpose."*

Analysis: This observation indicates that while AI can efficiently identify linguistic cues, its limited contextual awareness may reduce the perceived value of its feedback for fine-grained writing editing.

Excerpt 3: Preferred Feedback Source

DOI: https://doi.org/10.17509/ije.v18i2.85627
p- ISSN 1978-1342 e- ISSN 2442-4730

Interviewer: *"If you had to choose, would you employ more human response or AI feedback?"*

Student 3: *"For drafts, I'm fine with AI since it's quick. But for my final essay, I would rather have the teacher because it appears more personal and detailed."*

Analysis: The student shows a pragmatic approach to feedback sources, recognizing the complementary nature of human and AI input at different stages of the writing process.

Excerpt 4: Perceived Tone and Encouragement

Interviewer: *"Were there differences you noticed in the tone of the feedback by the AI compared to the teacher?"*

Student 4: *"Yes, definitely. The teacher's comments were more encouraging and personal. Even when they corrected errors, it was clear that they were attempting to get me to do better. ChatGPT's responses sometimes felt quite robotic or neutral — helpful, but not very inspiring."*

Analysis: This passage indicates how human feedback tends to have an affective component, offering encouragement that might raise student motivation. By contrast, AI feedback, as informative as it is, can be lacking this interpersonal warmth.

Excerpt 5: Feedback Specificity and Relevance

Interviewer: *"How relevant and precise did you consider feedback from the two sources?"*

Student 1: *"The instructor gave me examples from my own writing and suggested straightforward methods of improvement. The AI got the words right but didn't always point out how changing them would make my argument better. So, the instructor's feedback was more customized to my own requirements."*

Analysis: This response affirms the finding that teacher feedback is more context-sensitive and actionable, and thus, is interpreted by students as more pertinent to their individual writing concerns.

This study shows that *ChatGPT-4.5* has a relatively good ability to recognize hedges and boosters in student academic writing when compared to the evaluations of the human raters. This result is in agreement with Shalevska (2023) that AI-generated texts greatly displayed a high frequency of hedges which included, for example, the flexibility of "may" usage. However, these AI writing results, produced by *ChatGPT*, did not show the variety of usage or contextual appropriate usage by human writers. Further, Herbold et al. (2023) commented that although *ChatGPT*-produced essays were rated as higher quality, they often exhibited fewer discourse and epistemic markers than the human-written essays, suggesting a difference in pragmatic language use.

While there was overall agreement evidenced by the similar scores assigned by *ChatGPT* and the human teacher, other evaluative aspects diverged. The AI model produced students' feedback that was relatively generic and lacked the contextual relevance or didactic value that human evaluation provides overall. These two conclusions are in tune with Yoon et al. (2023), who reported that *ChatGPT* provided feedback that was abstract in nature and not particularly useful for English Language Learners, because it did not address specific suggestions for improvement. Shalevska (2023) also found that texts from AI models lacked boosters like "clearly" and "definitely" that are typical of human writerly expression and used as signals of certainty, which again highlights the limitations of AI in pragmatic expression.

The qualitative portion of student interviews revealed that students prefer human-generated feedback compared to those generated by AI with reasons stating that they are clearer and better-informing. These findings affirm that human feedback outperforms *ChatGPT*'s in clarity and pedagogical effectiveness. However, students recognized the worth of *ChatGPT* as a pre-writing assessment tool, in particular in relation to hedges and boosters, indicating that it might be useful as a complimentary tool in teaching writing.

Implications for AI Integration in Writing Assessment

The study's results indicate that while *ChatGPT-4.5* can accurately identify hedges and boosters, it does not provide the same appreciation and contextual knowledge with evaluative feedback as humans. This limitation emphasizes the importance of using AI tools with human input in an educational context. The recent emergence of AI in education creates a context within which schools, educators, and learners can act, weighing the potential of evolving technologies against irreplaceable human learning dimensions such as creativity and critical thinking.

4. CONCLUSION

This study contrasted *ChatGPT-4.5* performance with that of a human instructor in rating hedges and boosters in the writing of EFL students. Quantitative data showed that both raters produced statistically similar scores, and hence AI can be effectively employed to identify epistemic markers accurately. However, qualitative findings

DOI: https://doi.org/10.17509/ije.v18i2.85627
p- ISSN 1978-1342 e- ISSN 2442-4730

showed that the human teacher offered more richer pedagogical feedbacks, more responsive to rhetorical context and student intention. These results show that while AI offers precision and effectiveness in surface-level identification, it fails to offer the nuance needed for formative evaluation. The students themselves valued AI for first-stage feedback but chose human input in final drafts due to its sophistication and contextual richness.

Conceptually, the present study adds to the literature on AI-enabled evaluation of writing by raising the issues of the inability of current models to pragmatic and discourse-level evaluation. Future research would attempt to explore the integration of AI with annotated corpora in terms of rhetorical function and context and examine hybrid feedback models incorporating human and AI input with an aim to validating pedagogy in writing.

Future research is expected to analyze AI models trained on pragmatics and discourse annotated corpora with the goal of improving our evaluative capabilities. Moreover, research should examine the effect of combining AI-generated feedback along with teacher feedback on writing instruction. This, too, will provide more emphasis on the necessity of sequencing and combining AI and human feedback to achieve better student writing outcomes. In conclusion, while *ChatGPT-4.5* reflected great technical accuracy in hedges and booster detection, its response has no rhetorical delicacy and pedagogical sensitivity. These results corroborate the incorporation of AI tools as complements and not substitutes for human feedback in EFL writing instruction. The findings of this study thus contribute to the prism of knowledge on pragmatic competence in AI-enhanced pedagogy, warning against and pointing the way towards future development.

5. REFERENCES

Algaraady, J., & Mahyoob, M. (2023). ChatGPT's capabilities in spotting and analyzing writing errors experienced by EFL learners. *Arab World English Journal (AWEJ) Special Issue on CALL*, 9, 3-17. https://dx.doi.org/10.24093/awej/call9.1

Al-Mudhaffari, M., Hussin, S., & HoAbdullah, I. (2020). Interactional strategies in L2 writing: An exploration of Hedging and Boosting Strategies in Applied Linguistics research articles. *International Journal of Arabic-English Studies*, 20(1), 171–186. https://doi.org/10.33806/ijaes2000.20.1.9

Bok, E., & Cho, Y. (2023). Examining Korean EFL college students' experiences and perceptions of using ChatGPT as a writing revision tool. Journal of English Teaching Through Movies and Media, 24(4), 15–27. https://doi.org/10.16875/stem.2023.24.4.15

Charpentier-Jiménez, W. (2024). Evaluación de la inteligencia artificial y de la calibración de docentes en los cursos de escritura de inglés como lengua extranjera en una universidad pública costarricense [Evaluation of artificial intelligence and teacher calibration in English as a foreign language writing courses at a Costa Rican public university]. *Actualidades Investigativas en Educación*, 24(1), 1–25. https://doi.org/10.15517/aie.v24i1.55612

Dontcheva-Navratilova, O. (2016). Cross-Cultural Variation in the Use of Hedges and Boosters in Academic Discourse. *Prague Journal of English Studies*, 5(1), 2016. 163-184. https://doi.org/10.1515/pjes-2016-0009

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, *20*(1), 57. https://doi.org/10.1186/s41239-023-00425-2

Gayed, J. M., Carlon, M. K. J., Oriola, A. M., & Cross, J. S. (2022). Exploring an AI-based writing assistant's impact on English language learners. Computers and Education: Artificial Intelligence, 3, Article 100055. https://doi.org/10.1016/j.caeai.2022.100055

Geçkin, V., Kızıltaş, E., & Çınar, Ç. (2023). Assessing second-language academic writing: AI vs. Human raters. *Journal of Educational Technology and Online Learning*, *6*(4), 1096-1108. https://doi.org/10.31681/jetol.1336599

Herbold, S., Hautli-Janisz, A., Heuer, U., & Meurers, D. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, 13, 18617. https://doi.org/10.1038/s41598-023-45644-9

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. Continuum.

Karademir Coşkun, T. (2024). Unlocking the Future: The Role of Digital Learning Materials in Fostering 21st-Century Skills Among University Students. In R. Aggarwal, P. Gupta, S. Singh, & R. Bala (Eds.), *Augmented Reality and the Future of Education Technology* (pp. 229-252). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-3015-9.ch015

Kim, S., Shim, J., & Shim, J. (2023). A study on the utilization of OpenAI ChatGPT as a second language learning tool. Journal of Multimedia Information System, 10(1), 79–88. https://doi.org/10.33851/JMIS.2023.10.1.79

Ningrum, S., Puspita, H., & Mulyadi, A. I. (2024). Hedges and boosters in academic writing of ASEAN EFL learners. *Journal of English Education and Teaching*, *8*(1), 202–218. https://doi.org/10.33369/jeet.8.1.202-218EJournal Universitas Bengkulu

DOI: https://doi.org/10.17509/ije.v18i2.85627
p- ISSN 1978-1342 e- ISSN 2442-4730

Nguyen Thi Thu, H. (2023). EFL teachers' perspectives toward the use of ChatGPT in writing classes: A case study at Van Lang University. International Journal of Language Instruction, 2(3), 1–47. https://doi.org/10.54855/ijli.23231

Ryoo, S. M. (2023). Efficacy of hedges in formative feedback on L2 writing. *International Journal of Studies in Education*, 5(2), 550–567. https://doi.org/10.46328/ijonse.171

Shalevska, E. (2024). Hedges and boosters in AI and human writing: A comparative analysis. *Knowledge - International Journal*, *60*(1), 123–135. https://eprints.uklo.edu.mk/id/eprint/10825/

Shen, Z. (2023). Functions of hedges and boosters in academic writing. *Journal of Education Reform and Innovation*, 1(1), 13. (https://doi.org/10.61957/joerai-20230102)

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153–189. https://doi.org/10.3102/0034654307313795

Solak, E. (2024). Revolutionizing Language Learning: How ChatGPT and AI Are Changing the Way We Learn Languages. *International Journal of Technology in Education*, *7*(2), 353-372. https://doi.org/10.46328/ijte.732

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. Frontiers in Psychology, 14, Article 1260843. https://doi.org/10.3389/fpsyg.2023.1260843

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., ... & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, *91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Tran, T. S., Nguyen, S. D., Lee, H. J., & Tran, V. P. (2023). Advanced crack detection and segmentation on bridge decks using deep learning. *Construction and Building Materials, 400*, Article 132839. https://doi.org/10.1016/j.conbuildmat.2023.132839

Wingate, U. (2014). The impact of formative feedback on the development of academic writing. In *Approaches to Assessment that Enhance Learning in Higher Education* (pp. 29-43). Routledge.

Xiao, Y., & Zhi, Y. (2023). An exploratory study of EFL learners' use of ChatGPT for language learning tasks: Experience and perceptions. Languages, 8(3), Article 212. https://doi.org/10.3390/languages8030212

Xiao, F. (肖斐文), Zhu, S. (朱思宇), & Xin, W. (辛闻). (2025). Exploring the landscape of generative AI (*ChatGPT*)-powered writing instruction in English as a foreign language education: A scoping review. *ECNU Review of Education*, 0(0). https://doi.org/10.1177/20965311241310881

Yoon, S.-Y., Miszoglad, E., & Pierce, L. R. (2023). Evaluation of ChatGPT feedback on ELL writers' coherence and cohesion. *arXiv preprint arXiv:2310.06505*. https://arxiv.org/abs/2310.06505

DOI: https://doi.org/10.17509/ije.v18i2.85627
p- ISSN 1978-1342 e- ISSN 2442-4730