AI, Human, and Hybrid Approaches to Language Assessment: Analyzing EFL Writing across Cognitive and Contextual Dimensions

Sohaib Alam

DOI: 10.18355/XL.2026.19.04.01

Abstract
The increasing integration of artificial intelligence (AI) into language assessment has intensified the need to examine its effectiveness in evaluating both the cognitive and contextual aspects of second language writing. This study compared AI-only, human-only, and hybrid AI-human assessment of English as a Foreign Language (EFL) argumentative writing across three cognitive dimensions (logical organization, argumentative reasoning, and idea development) and three contextual dimensions (contextual appropriateness of language, audience awareness, and register and tone appropriateness). A quantitative, comparative, cross-sectional research design was employed using 167 argumentative essays produced by undergraduate EFL students at a private university in Saudi Arabia. Experienced human examiners and ChatGPT (GPT-4o) independently evaluated the essays using an identical analytic rubric, and hybrid scores were derived by averaging the human and AI evaluations. Descriptive statistics, intraclass correlation coefficients, weighted kappa coefficients, repeated-measures analysis of variance, and Bonferroni-adjusted pairwise comparisons were conducted to compare evaluation outcomes, agreement, and score differences. The findings demonstrated broadly comparable scoring patterns across both cognitive and contextual aspects of writing. The three assessment methods showed moderate to good agreement, though agreement was lower for contextual dimensions requiring more interpretive judgment. Statistically significant differences were identified for only three dimensions, with uniformly small effect sizes indicating limited practical significance. Overall, the findings suggest that AI can effectively complement human judgment under standardized assessment conditions, while hybrid assessment offers a balanced approach that integrates computational consistency with human interpretive expertise for multidimensional EFL writing assessment.
Key words: Artificial intelligence, ChatGPT, EFL writing assessment, hybrid assessment, human assessment, technology-enhanced language assessment, SDG-4, educational technology

1. Introduction

The validity of English as a Foreign Language (EFL) writing assessment depends on how assessors evaluate measurable linguistic aspects as well as higher-order communicative aspects, such as argumentative reasoning, audience awareness, and appropriate language use within the writing context. The multi-dimensional character of writing is important for assessment both from a pedagogical and methodological point of view. This multi-dimensionality reveals the tension inherent in current practice: the need for efficiency in assessing large numbers of students, the need for interpretation and complexity found in valid assessment. This tension has grown with the upsurge of AI-driven writing assessment systems, which bring remarkable efficiency and uniformity but also speak to unanswered questions regarding their

ability to capture the construct of writing proficiency (Li et al., 2024; Bouziane & Bouziane, 2024).

Recent research confirms that AI systems can evaluate surface-level linguistic features such as grammatical accuracy, lexical sophistication, and textual organization with considerable consistency across large volumes of student writing (Li et al., 2024; Tran et al., 2025). However, their ability to assess higher-order dimensions, including logical organization, idea development, and register appropriateness, remains empirically contested. Studies reporting limitations in AI evaluation of argumentative reasoning and audience awareness suggest that automated assessment alone may not provide a sufficiently valid representation of writing proficiency (Lo et al., 2025). At the same time, exclusive reliance on human assessment introduces its own difficulties: rater variability, inconsistent application of scoring criteria, workload constraints, and limited scalability in higher education settings (Weigle, 2002; Eckes, 2012). Neither approach, taken independently, appears adequate for the demands of contemporary EFL writing assessment.

As a response, hybrid AI-Human assessment has been suggested as an integrated evaluative framework that combines the analytical efficiency of AI-based systems with the context-oriented judgment, disciplinary expertise, and holistic evaluation of human evaluators (Ding et al., 2025; Khojasteh et al., 2025; Alam, 2026). One method is for AI to examine objective linguistic and structural elements, while human assessors evaluate rhetorical qualities that require communicative interpretation. While this approach rests on a widely established conceptual framework, researchers have yet to agree on the appropriate implementation method; studies have oscillated between under-theorized dichotomous integration and weighted integration. More importantly, few studies have systematically compared the performance of AI-only, human-only, and hybrid AI-human writing while evaluating the same cognitive and contextual dimensions of writing within the same assessment model. Consequently, little empirical evidence is available to clarify how strong or weak these assessment methods are, and when and how they can be appropriately used to evaluate each aspect of EFL writing.

To tackle this issue, the present study compares three approaches to assessment (AI only, human only, and hybrid AI-human) for six dimensions of EFL writing: logical organization, argumentative reasoning, and idea development (cognitive dimensions), contextual appropriateness of language, audience awareness, and register and tone appropriateness (contextual dimensions). This dimensional structure rests on three complementary theoretical viewpoints. The cognitive dimensions examined here are based on cognitive models of writing which were developed by Flower & Hayes (1981) and Bereiter & Scardamalia (1987), that see writing as a recursive process of planning, organizing, and revising. This introduces the three contextual dimensions by using the perspective of systemic functional linguistics (SFL) (Halliday, 1985) which is a frame of reference that views language as a tool for creating meaning, with its success dependent on the audience, purpose, and social context in which it is used. The framework of Technology-Enhanced Language Assessment (TELA) (Shermis & Burstein, 2013; Warschauer & Grimes, 2008) not only serves as one approach to assessing how technology can and should be used, but it assesses what traditional, human where technology alone is insufficient. These perspectives should be read

together for the benefit of a principled assessment in relation to different modes of assessment, not to endorse any particular approach by itself.

Researcher has formulated three research questions to guide the study. First, how do assessments of the cognitive and contextual aspects of EFL writing differ between the AI-only, the human-only, and the hybrid AI-human's assessments? Second, what are the inter-rater reliability indices of the three approaches, indicating that there is some level of agreement and consistency of scoring? Third, do the mean scores for the three approaches differ significantly on the six dimensions of writing? The paper fills a gap in the empirical research of technology-enhanced language assessment and offers helpful advice for teachers and institutions using AI in the assessment of writing in higher education.

2. Literature review

2.1 Human Assessment of Writing

Human assessment has always played a central role in the study of writing assessment; it is widely endorsed for its ability to consider written texts in relation to their linguistic, rhetorical, and communicative contexts. The linguistic aspect is complemented by human evaluation, which assesses students' individual characteristics as reflected in their texts, including their ideas, arguments, audience awareness, and the appropriateness of language to the communicative context. That is why, until now, human judgment has been the criterion of most constructive validity when evaluating writing and is the framework against which new technological tools are measured (Weigle, 2002; Eckes, 2012). There are cautions to the benchmark status, however; the reliability of human scores is far from being unquestionable, especially in large-scale settings where differences between scores may result in less accurate judgments of scores and decisions about them (Lumley, 2002; Knoch, 2009). Due to the hermeneutic depth and subjectivity of human judgement, alternative and complementary assessment methods have been extensively studied within the last few years.

Empirical research consistently demonstrates that human assessment is particularly effective in evaluating complex writing constructs that extend beyond measurable linguistic features. Experienced raters have been shown to identify subtle strengths and weaknesses in learners' texts, demonstrating sensitivity to rhetorical effectiveness, communicative intent, and idea development that automated systems do not consistently replicate (Bouziane & Bouziane, 2024; Li et al., 2025). In addition, human-generated feedback has been found to provide richer, more context-sensitive evaluations that contribute more substantially to learners' writing development than automated feedback alone (Jovic et al., 2025; Lo et al., 2025). These advantages are particularly evident when assessment targets higher-order cognitive dimensions such as the coherence of an argument or the appropriateness of register for a specific audience where evaluative judgment depends on contextual knowledge that goes beyond pattern recognition.

However, several practical and methodological issues have been documented, highlighting the continued importance of human assessment. Probably the most recurrent concern has been one of rater variability, which occurs because of rater differences in professional expertise, interpretation of scoring guidelines, and scoring

tendencies. Importantly, the variability observed varies in type and size with the type of scoring used. In holistic assessment, the raters combine more than one quality into one score, and where there are systematic differences in ratings for some aspect of the writing, they may cover up. There can be more or less variance in analytic assessment, where each assessment criterion is scored separately, and variability may occur across different scoring criteria. For example, the variance in scores might be greater across dimensions of audience awareness, reasoning, and organization than across dimensions of organizational structure and/or contextual and rhetorical awareness by a rater (Knoch, 2009; Eckes, 2012). As the current study uses percentile, analytic scores in each dimension, across the six cognitive/ contextual categories, an understanding of human rater variability in each of the dimensions is directly relevant when interpreting the findings. In addition to the rater variability, human evaluation is extremely time consuming, restricting scalability and impeding the ability to provide timely feedback, especially in large numbers, particularly in the higher education environment. These constraints have led researchers to examine the potential of technology-based alternatives to improve the efficiency of assessment while maintaining the interpretive benefits of human expertise.

2.2 AI-Based Writing Assessment

With the advent of AI-based writing assessment, the scope of writing assessment has greatly expanded in EFL contexts. Through the use of these technologies, including NLP, machine learning and large language models, contemporary AI systems can analyse written texts across various facets of language and discourse. Recent development in NLP, machine learning and large language models have helped present AI system to increase their understanding of smaller aspects of written language and discourse. Some LLM based systems do more than just errors at the surface of the text, they recognise other aspects of writing, including the overall coherence, organization of text, vocabulary use, and communicative effectiveness of the writing (Fathi & Rahimi, 2024). The systems that have garnered the most scholarly attention in the field of EFL assessment are the ones that were developed and tested as assessment systems: ChatGPT for its ability to process extended written discourse; interpret rubric criteria, and generate assessment feedback; for its capacity to write assessment commentary on a multitude of dimensions of written quality (Alam et al., 2026). For this reason, the AI assessment tool which is chosen for the present study is the ChatGPT system. ChatGPT has been utilized in other comparative research in the EFL field to assess EFL learners' writing in various analytic aspects of the genre (e.g., Du & Nordin, 2025; Li et al., 2024).

AI has been demonstrated to improve assessments in terms of effectiveness, standardization, and scalability. The benefits of AI (assessment) highlights efficiencies, uniformity and scalability. Unlike human scorers, who can become fatigued, feel emotions fade and drift in context over time, AI can rapidly scan a huge volume of student writing, apply the concepts of a scoring rule, and do it all without getting tired, losing motivation, or misinterpreting what a score means (Du & Nordin, 2025; Li et al., 2024; Tran et al., 2025). Of the conclusions drawn in the studies based on the characteristics of performance: grammatical or otherwise, vocabulary good or bad, or the organization of text, there was high coherence across studies. Given all that, it is not difficult to imagine that the aspects of utility have led to an increasing institutional reluctance to rely on humans for assessment services, typically in the

background, as well as an alternative option to humans, in large-scale education projects.

AI assessment still has a number of disadvantages (Kobylarek & Birova, 2026), and are more applicable to the context of writing assessment, focusing specifically on what is going on in the moment of the writing task. Recent studies show that AI systems perform less well on assessments and grading higher-order constructs that require a researcher's interpretation beyond accuracy, rhetorical elements and context; what dimensions have been identified to define proficient academic writing for assessment? (Hartwell & Aull, 2023; Lu et al., 2024) Furthermore, Bouziane & Bouziane (2024) conducted a study showing that AI use to assess argumentative reasoning and audience awareness was predictive to a lesser extent than grammatical correctness and vocabulary. This could introduce systematic bias between AI evaluations and expert human evaluations in how they cover constructs. Lo et al., 2025 found similar patterns in their study of the AI systems: when the writers deliberately and skillfully chose a style that departed from traditional academic writing, but was rhetorically suitable, the systems were unable to perform effectively. These results suggest that questions regarding the validity of assessment using AI should be raised, among others, on the cognitive and contextual sub-skills covered by the present study, as this only partly seems to be reflected in the AI-only assessment format. Furthermore, evaluating writing processes using AI is problematic from a construct-validity perspective, since the linguistic characteristics that identify non-native varieties of English are automatically treated as “wrong” language use, while linguistic gaps are interpreted as “deficiency” in the English language. Another concern is algorithmic bias with regard to the use of artificial intelligence in the assessment of the writing process, especially the automatic detection and labeling of linguistic features that are recognized as non-native varieties of the English language as “wrong” and any gaps as “deficiency” of the English language or variety. These indications suggest that although AI assessment is proven effective and standardized, it doesn't replace human interpretation and expertise.

2.3 Hybrid AI-Human Assessment

The study of hybrid evaluation, where human and AI assessment are used together in the same evaluative framework, has been driven by the recognition that each has strengths and weaknesses. In this study, hybrid AI-human assessment refers to an assessment process that combines AI scores with human rater scores in an integrated manner within the context of evaluative scores that are evaluated by both modalities using all distinct writing dimensions by a simple averaging of the scores assigned independently by each of the modalities for each writing dimension. This aligns with hybrid designs recently demonstrated in empirical studies that use AI and human evaluations as complementary, not competitive, sources of evidence (Ding et al., 2025; Khojasteh et al., 2025). The idea is that AI systems can be used to assess structural and linguistic aspects with objectivity and consistency, whereas human raters can provide the structural, linguistic, and supporting context and rhetorical insight that is required for evaluative judgment on the dimensions of writing at higher levels. A hybrid assessment means that such contributions are being brought together to get a more comprehensive coverage of the construct than each approach on its own merits (Kobylarek, 2026).

The research evidence base is diverse and just beginning to emerge on hybrid assessment, and the evidence needs to be interpreted with caution. Ding et al., (2025) examined the use of AI, humans, and hybrid (combined) assessments to evaluate ACW and found that the hybrid approach unsurprisingly outperformed the other two approaches on the external validity criteria, with some of these differences on the external validity criteria that require storing and making contextual judgments. Similarly, Khojasteh et al., (2025) revealed that hybrid assessment led to both more consistent and reliable assessments of student writing and to fewer individual rater bias effects, while still being sensitive to higher-order writing skills. Asadi et al., (2025) suggested that the answers provided by the AI writing tools were clearer and enhanced the instructional value of written assessments when supported by teacher feedback. In conclusion, the findings suggest that the calculation will extend construct testing by combining the particles' ability to assess logical correctness with the fact that some particles will be of human origin. However, the integration process varied from study to study, ranging from human raters scoring the AI's scores before humans' actual final judgments, to parallel independent scoring approaches, and then averaging scores, to an adjudicating process for scores, to more iterative processes where the humans consulted along with the AI's analysis. Results from different design characteristics make it hard to compare results across studies and underscore the need to report how the study has been transformed, as done in the current study.

Yet, plenty of issues are yet to be resolved in order for hybrid assessment to be implemented and evaluated. Several unanswered questions remain regarding the implementation and evaluation of hybrid assessment. First, there is no agreement on the specific ratio of AI-generated to human judgment, and it remains unclear whether these two types of scores are equally weighted across writing dimensions or whether some dimensions are better served by relying more on one type of input. Second, the extent to which hybrid scoring decisions align independently with either AI or human judgments in contexts with significant disagreement has not been studied systematically: most studies report only average-level results and focus on dimensions and/or score ranges with smaller disagreements. Third, although most studies in recent years conclude that hybrid assessment performs better than either full-level assessment or item analysis alone, it is significant that this conclusion is based almost entirely on studies that reported positive results; this could be because the summary of results is a product of publication bias, not the actual diversity of hybrid assessment results. Publications of studies which found no benefits from integration over human assessment were relatively rare, and some studies added to human assessment inconsistencies that didn't exist in the original work. These constraints suggest that hybrid assessment should be interpreted with caution and evaluated empirically, across in multiple dimensions of writing and in a few contexts of assessment, which the present study aims to examine.

2.4 Theoretical Framework

The present study is guided by three complementary theories that posit theories about how the writing process works, which are used as the basis for deriving the analytical framework: Cognitive Models of Writing, SFL, and Technology-Enhanced Language Assessment. The lenses presented here are not treated as completely distinct theories per se, but as complementary windows that can help explain different aspects of

writing quality, both cognitive and contextual, addressed in this study. Table 1 maps the theoretical perspectives to the writing dimensions they support.

Cognitive Models of Writing view writing as a recursive and iterative goal-directed process in which writers plan/cooperatively/narrate and then review/revise their writing (Flower & Hayes, 1981). A main point of this model is that writing is a problem-solving process that requires writers to make content, rhetorical, and linguistic decisions simultaneously. In addition, Bereiter and Scardamalia (1987) expanded this schema by separating knowledge-telling (recalling content in sequence) from knowledge-transforming (restructuring content to communicate and/or make it argumentative). In this way, the quality of a piece of writing is most evident in how well a writer has decomposed their ideas into a logical sequence, argued for them, and made content-based decisions. Logical organization, argumentative reasoning, and idea development, therefore, are the cognitive dimensions studied in the present study, since they encompass the cognitive functions (strategic planning, problem-solving, and knowledge-transforming capacities), attributed to the models. Although the Flower-Hayes and Bereiter-Scardamalia theories omit the sociocultural and collaborative aspects of writing emphasized by more recent scholarship (Prior, 2006), they remain relevant for describing the cognitive processes required to create high-quality text, which this study assesses.

In fact, the systemic functionalist theory of language (SFL) (Halliday, 1985) regards language not as a system of formal rules, but as a semiotic resource for meaning-making in social and communicative situations. A key concept in SFL is the notion of the register, which refers to the systematic variation in the use of language across the three register variables of field ("what happens" and "what is it about?"), tenor ("what is being done" and "what is it becoming?") and mode ("how much formality" or "which medium"). These three variables directly underlie the three contextual dimensions we studied in the present study. The appropriateness of language as a quality of writing is reflected by writers' sensitivity to the field; that is, by how language choices turn out to be appropriate to the academic content and expectations of the task. Audiences indicate tenor or the level to which writers attend to the reader's expectations, knowledge and communicative needs. Correctness and/or tone appropriateness and communicative stance refer to maintaining a level of formality and communicative stance that aligns with the conventions of academic writing. The present study makes these dimensions meaningful through the SFL framework rather than as stylistic preferences, treating them as dimensions of communicative competence that reflect writers' understanding of how language operates in context.

The researcher proposes to examine how AI tools fit with existing human assessment tools as complementary rather than competing; the researcher adopts the Technology-Enhanced Language Assessment framework to understand this (Shermis & Burstein, 2013). In this view, automatic systems would be effective analysis devices that detect structures in natural language, and human evaluators would be agents of interpretation who would impose meaning on the linguistic phenomena in their communicative and disciplinary context. New research in this area has focused on the necessity for the validity of the technology-mediated score to represent the construct it is designed to measure, as well as the conditions under which it requires human input to ensure the assessment scores are aligned with the complexity of writing proficiency (Chapelle & Voss, 2016; Deane, 2013). These are directly relevant to the comparative framework

of the present study, which includes an assessment of the construct validity of AI-only, human-only and hybrid assessments across dimensions that differ in their capacity for automated analysis.

Table 1. Theoretical Framework and Dimension Mapping

Theoretical Perspective	Core Construct	Writing Dimensions in This Study
Cognitive Models of Writing	Recursive planning, knowledge transformation, argumentative problem-solving	Logical organization; Argumentative reasoning; Idea development
Systemic Functional Linguistics	Register (field, tenor, mode); language as contextual meaning-making	Contextual appropriateness of language; Audience awareness; Register and tone appropriateness
Technology-Enhanced Language Assessment	AI as analytical tool; human as interpretive agent; construct validity in automated assessment	Framework for comparing AI-only, human-only, and hybrid approaches

2.5 Research Gap

From the above review, it is clear that studies in the literature have been conducted as distinct research programs and assessed on their own merits, without head-to-head comparisons across assessment types. As a result, educators and institutions today lack empirical evidence to inform decisions about the best method (or methods) for assessing the development of particular aspects of writing quality. The observation is followed by three interrelated gaps.

The first and most important mismatch is the lack of an explicit, in-framework comparison of all three assessment modalities. An approach to studying both the human and AI baseline is generally not used in studies that examine hybrid assessment, but rather a single baseline is first studied. The most similar previous studies to the current ones are those conducted by Ding et al. (2025) and Khojasteh et al. (2025), which compare the results of AI-assisted and human evaluation approaches, but neither has a fully independent hybrid condition in which the scores averaged from both the AI and the human evaluators are compared against the score of the AI alone or the human alone, across the same number of analytic dimensions. Without this three-way comparison, it is impossible to discern whether hybrid assessment provides evaluative value beyond what each approach provides.

The second one relates to the dimensions of the current comparisons being made. Surprisingly, past studies have mainly concentrated on overall writing quality results or on linguistic aspects more visible to the observer, such as accuracy in grammatical structures, among others. In comparison to the attention that has been paid to evaluating the performance of the various approaches in higher-order cognitive dimensions logical organization, argumentative reasoning, and development of ideas relatively little attention has been dedicated to assessing the performance of approaches with respect to assessing contextual dimensions: one's (counter) particular

audience awareness, register appropriateness and contextual appropriateness of language. This is a major mistake, as those dimensions are exactly where the major simplifications of AI have been cited and where the possibility for human judgment to complement AI is theoretically most attractive.

The third gap concerns measure of agreement and consistency between different assessment approaches. Few studies have examined how well two types of AI (AI only vs. human only) and a hybrid version (AI-defined vs. human-defined) align on the same texts with the same analytic traits. How these modalities converge, or not, across cognitive or contextual dimensions remains largely unknown, as does whether this convergence is consistent across dimensions. This divergence gap matters because high divergence on some dimensions would suggest they measure different constructs, with implications for the validity of using one modality over another for a given dimension. The aim of the present study is to address these three gaps: three-way comparison, consideration of higher-order cognitive and contextual aspects and the lack of supporting evidence on cross-modal agreement.

2.6 Present Study

To address the identified gaps, the present study compares AI-only, human-only, and hybrid AI-human assessment of EFL argumentative essays within a single evaluative framework. In the AI-only condition, ChatGPT evaluates each essay independently using a purpose-developed analytic rubric. In the human-only condition, trained EFL writing instructors evaluate the same essays using the same rubric. In the hybrid condition, final scores are calculated by averaging the AI-generated and human-assigned scores for each dimension. All three conditions are applied to the same set of student essays and evaluated across the same six dimensions, enabling direct comparison of assessment outcomes, agreement levels, and score differences. The findings are expected to provide empirical evidence on the relative validity of each approach and to inform theoretically grounded decisions about integrating AI and human judgment in EFL writing assessment in higher education. The study is guided by the following research questions (RQ):

RQ1: How do AI-only, human-only, and hybrid AI-human assessment approaches differ in the pattern and direction of scores assigned across the cognitive dimensions (logical organization, argumentative reasoning, idea development) and contextual dimensions (contextual appropriateness of language, audience awareness, register and tone appropriateness) of EFL writing?

RQ2: To what extent do AI-only, human-only, and hybrid AI-human assessment approaches demonstrate agreement and consistency in evaluating EFL writing, as measured by intraclass correlation coefficients and weighted kappa indices across the six analytic dimensions?

RQ3: Are there statistically significant differences in mean scores among AI-only, human-only, and hybrid AI-human assessment approaches across the six cognitive and contextual dimensions of EFL writing?

3. Research methodology

3.1 Research Design

The current paper adopts a quantitative comparative cross-sectional research design to investigate differences in evaluation results, score agreement, and score distribution among three kinds of assessments, namely, human-only assessment, AI-only assessment, and hybrid assessment (combining AI with Human assessment). A quantitative design was chosen for this study since it was used to compare the scores of the three assessment methods, to examine the level of agreement and consistency between the assessment methods, and to determine if there was a statistically significant difference among the six writing dimensions. An identical analytic scoring rubric was used to independently evaluate the same set of argumentative essays, facilitating comparability across assessment conditions. This design helped systematically generate numerical data and apply inferential statistical analysis to answer the study's research questions (Creswell & Creswell, 2017).

3.2 Research Context

The study was carried out at a public university in Saudi Arabia where English as a foreign language is compulsory for all undergraduate programs. In addition, some of the major subjects are taught in English, many of which involve academic writing. The study was set in undergraduate English writing courses where argumentative essay writing is a central topic area and a regular assessment task. This educational context, in turn, was deemed suitable for comparing human-only, AI-only, and hybrid (AI-human) assessment methods for analyzing the cognitive and contextual aspects of EFL writing.

3.3 Participants

The subjects consisted of 167 undergraduate EFL students (in the English Writing class) studying at the study institution. Because participants were easily accessible and available, and because argumentative writing was a part of coursework itself, a convenience sampling technique was used in this study (Cohen et al., 2018; Cheng & Dörnyei, 2007). We performed an a priori power analysis using G*Power 3.1 (Faul et al., 2007) to calculate the minimum sample size required for the planned statistical analyses. If the effect size is medium (f = .25, Cohen, 1988), that analysis revealed a sample of at least 128 participants for a level of statistical power of .80 and at a significance level of $\alpha = .05$. The final sample of 167 participants was sufficiently adequate for the statistical power to show a statistically significant difference between the three assessment approaches.

The sample included 94 male (56.3%) and 73 female (43.7%) students aged 18 to 23 years (M = 20.41, SD = 1.26). The participants, based on the placement procedure used at the university, had a general intermediate to upper intermediate level of English language skill (CEFR B1-B2) (Council of Europe, 2001), with 89 students (53.3%) at B1 level and 78 students (46.7%) at B2 level. To include a teacher in this study, they needed to be an official participant in the specified English writing course, and they had to complete the argumentative writing assignment in a standardized, classroom-based environment and submit essays that were long enough and comprehensive enough for analytic assessment. Essays not meeting the above

requirements were not included in the corpus, and the corpus consisted of 167 argumentative essays for analysis.

3.4 Writing Task

The data comprised 167 argumentative essays written under standardized classroom conditions. All participants responded to the same writing prompt:

"Some people believe that universities should focus primarily on providing students with practical skills for employment, while others argue that universities should prioritize the development of critical thinking and intellectual inquiry, regardless of immediate career relevance. Discuss both views and give your own opinion, supporting your argument with relevant reasons and examples."

The prompt was selected because it required students to evaluate contrasting viewpoints, construct and defend a position, and support their arguments with relevant reasons and examples. These task requirements elicited performance across the six writing dimensions examined in the study: logical organization, argumentative reasoning, idea development, contextual appropriateness of language, audience awareness, and register and tone appropriateness (Bereiter & Scardamalia, 1987; Halliday, 1985; Knoch, 2009; Weigle, 2002). Participants were allocated 60 minutes to complete the task and were instructed to write approximately 250-300 words without access to dictionaries, internet resources, or AI-assisted writing tools. The time allocation reflected the institution's standard assessment policy and ensured that all participants completed the task under identical testing conditions.

3.5 Assessment Framework

Three assessment methods (human, AI, and hybrid AI-human) were implemented on the same set of argumentative texts. All three approaches used the same analytic scoring rubric of six dimensions (logical organization, argumentative reasoning, idea development, contextual appropriateness of language, audience awareness, and register and tone appropriateness) to facilitate measurement across the three assessment conditions. The employment of an analytic rubric was found to be appropriate, as it would allow for an evaluation of the writing process at each dimension and allow for direct comparisons of assessment outcomes across the dimensions, across in-service sessions and across various assessment modalities (Knoch, 2009; Eckes, 2012).

The human-only assessment was carried out by two experienced EFL writing teachers who both have a Ph.D. in Applied Linguistics or TESOL, and have over five years of experience in English as a foreign language writing, with one whose background is in English language teaching and the other in assessment. Raters went through a norming process before the assessment, during which they developed a shared understanding of the concepts in the rubric for analyzing and scoring the assessment items, and to ensure that one rubric was applied equally across all students. Each rater then independently rated all the essays, blinded to the other rater's ratings and the AI-generated ratings. When differences were greater than one scale point on each of the writing dimensions, the differences were discussed until the students reached a consensus. In doing so, both raters re-read the scoring rubric and the essay under consideration, and then came to a consensus about a final score, which became the

human assessment score for that dimension. (Weigle, 2002; Lumley, 2002; Eckes, 2012).

This assessment was conducted using AI only with ChatGPT (GPT-4o, OpenAI). Individually, each essay was rated on a standard outline that contained the full analytic rating scale and asked the model to rate the paper on the six dimensions of writing on a scale of 1 through 5. Consistently applying the same version of the ChatGPT model, the same evaluation prompt, and the same analytic grading guide were undertaken to ensure that the conditions of evaluation and comparison of all writing samples were the same. All essays were handled in the same way in terms of the prompt and all the essays were evaluated under the same scoring criteria to ensure evaluation conditions are standardized between essays (Bouziane & Bouziane, 2024; Du & Nordin, 2025; Li et al., 2024). Both the AI-based and the human-based evaluations were made via separate processes and without any reference to each other.

The hybrid, AI-human assessment is created by merging the consensus scores from the human assessment with the AI-generated scores from ChatGPT. The final hybrid writing score was the mean of the human consensus score and the AI score, as in previous studies that have explored hybrid assessment models (Asadi et al., 2025; Ding et al., 2025; Khojasteh et al., 2025). An equal-weight integration system was used to ensure that the data from the AI and human evaluations could be incorporated into the seamless integration without any one source being privileged over the other, based on the absence of any reliable empirical evidence to support prioritizing one data source over another. This study adopted this hybrid assessment as an equally embedded assessment method for comparison purposes, not as a method to conclude which is the best way to integrate AI and Human scores. Using the same writing samples, scoring rubric, and assessment criteria across all three approaches ensured consistency in the assessment process and allowed direct comparisons of assessment outcomes.

3.6 Assessment Instrument

The present study developed an analytic scoring rubric to assess EFL writing in terms of 6 dimensions that come from the study's theoretical framework. The rubric included three cognitive dimensions that were based on Cognitive Models of Writing (Flower & Hayes, 1981; Bereiter & Scardamalia, 1987): logical organization, argumentative reasoning, and idea development; and three contextual dimensions, based on Systemic Functional Linguistics (Halliday, 1985): contextual appropriateness of language, audience awareness, and register and tone appropriateness. The five-point rating scale was used to assess each dimension separately, and higher scores indicated superior writing performance. The analytic format was chosen, as it allows assessment of writing quality on a dimensional level and comparison of the writing assessment performance with the results of the human-only, AI-only, and hybrid AI-human writing assessment approaches (Knoch, 2009; Eckes, 2012).

The rubric was approved by a group of five EFL writing assessment and Applied Linguistics experts with long experience in writing assessment and rubric construction, in order to gain content validity. The experts reviewed the clarity of the scoring descriptors, along with how well they were relevant and representative, and revisions were made until there was “consensus” on the instrument's sufficiency

(Weigle, 2002; Pill & Harding, 2013). The rubric was then tested with ten argumentative essays not collected as part of the main data set and was subsequently reviewed by experts. The results of the pilot study showed that the six dimensions of assessment were well discriminated between using the five-point continuum and the descriptors were generally easy to interpret. The final analytic scoring rubric was then used for all scores administered by humans, AI, and hybrid AI-human. The full analytic grading guides are in Appendix A and the uniform assessment prompt for using ChatGPT with artificial intelligence is in Appendix B.

3.7 Data Collection Procedure

The procedure of data collection was consistent throughout the study. The participants wrote their arguments first in a classroom setting with supervision. Upon completion, all essays were gathered and anonymized by giving each individual writing sample an identifying code before the assessment process started. The essays were then judged using each of the three assessment strategies in an anonymous manner. The AI-only assessment was performed after the human-only assessment and the standardized ChatGPT evaluation prompt. Once both assessments were done separately, hybrid AI-human assessment scores were computed by fusing the consensus scores from the humans with the scores from the AIs as explained in Section 3.5. Lastly, all assessment scores were coded into a structured database, reviewed, and double-checked for accuracy and completeness.

3.8 Data Analysis

The data were analyzed by IBM SPSS Statistics (Version 31). The analyses were done according to the three hypotheses formulated in the study. Descriptive statistics of the mean, standard deviation, minimum scores, and maximum scores for each assessment approach (six writing dimensions, Field, 2018) were obtained for each writing approach. To address RQ2, the level of agreement and consistency among the human-only, AI-only, and hybrid AI-human assessment approaches was examined using the Intraclass Correlation Coefficient (ICC) based on a two-way mixed-effects model with absolute agreement (Shrout & Fleiss, 1979; Koo & Li, 2016), while weighted kappa coefficients were computed to assess pairwise agreement for each writing dimension (Cohen, 1968; Landis & Koch, 1977). The distribution of the data was checked using the Shapiro-Wilk test and by examining skewness, kurtosis, and Q-Q plots (Field, 2018; Tabachnick & Fidell, 2019) prior to carrying out any analyses of the data with inferences. For the third research question, we analyzed differences between the three approaches to assessing the variables using Repeated-Measures Analysis of Variance (RM ANOVA), with Greenhouse-Geisser corrections when the assumption of sphericity was violated. The Friedman test was used for variables where parametric test assumptions could not be met as a non-parametric alternative to the Friedman test. If statistically significant differences were found, Bonferroni-corrected pairwise comparisons were used to identify the between-group differences. A p-level cutoff of $< .05$ was used for statistical significance, and partial eta-squared values (η^2p) and Cohen's d values were reported as a measure of the magnitude of the observed effects (Cohen, 1988).

4. Results

The assessment scores were then analyzed to check that the statistical analyses planned were suitable for the data. As would be expected in a relatively large sample size (N = 167) a few of the variables showed statistically significant departures from normality ($p < .05$), as normality tests are sensitive to small deviations in larger sample sizes. Given this, the normality assumption was assessed by considering several complementary measures instead of just the significance of the Shapiro-Wilk test. The skewness, kurtosis and normal Q-Q plots were used to confirm that the distribution were approximately normal, with values for skewness and kurtosis within range of ±1.5. Thus, it was assumed that the assumptions in the parametric analyses were adequately met. Mauchly's test showed that for the dimensions Audience Awareness and Register and Tone Appropriateness, the assumption of sphericity was not met, and Greenhouse-Geisser corrections were performed for these variables in the repeated-measures analyses. Repeated measures analysis of variance (ANOVA) was therefore used as the main inferential statistics. The Friedman test was followed as a confirmatory non-parametric test and the findings were found to be consistent with the findings arising from repeated measures ANOVA. This is because both analytical solutions produced the same general pattern of findings which is why just the parametric results are reported.

The first research question focused on distinguishing the differences among the assessment modes (AI-only evaluation, human-only evaluation and hybrid (AI/human) evaluation) in addressing the cognitive and contextual aspects of EFL writing. The descriptive data of the six writing dimensions are presented in Table 2.

Table 2. Descriptive Statistics for Human-only, AI-only, and Hybrid AI-Human Assessment

Writing Dimension	Assessment Approach	Mean	SD	Minimum	Maximum
Logical Organization	Human	3.63	0.62	2	5
	AI	3.76	0.58	2	5
	Hybrid	3.70	0.55	2	5
Argumentative Reasoning	Human	3.42	0.68	2	5
	AI	3.36	0.64	2	5
	Hybrid	3.39	0.60	2	5
Idea Development	Human	3.48	0.66	2	5
	AI	3.41	0.61	2	5
	Hybrid	3.45	0.58	2	5
Contextual Appropriateness of Language	Human	3.58	0.60	2	5
	AI	3.51	0.57	2	5
	Hybrid	3.55	0.54	2	5
Audience Awareness	Human	3.46	0.65	2	5
	AI	3.27	0.63	2	5
	Hybrid	3.37	0.59	2	5
Register and Tone Appropriateness	Human	3.61	0.59	2	5
	AI	3.44	0.61	2	5
	Hybrid	3.53	0.56	2	5

As shown in Table 2, the rubrics produced very similar ratings across all six writing assessment dimensions for the three assessment methods. Across the range of scores, mean scores ranged from 3.27 to 3.76, indicating comparable overall levels of students' writing ability regardless of assessment method. Small differences were noted at each dimension, but were held consistently small among the three

approaches. Compared with the human-only assessment, the largest difference was in Audience Awareness (+/.09 points), with the others being minor.

The AI-only assessment was more consistent with the human raters' scores than the other two ratings for Audience Awareness and Register and Tone Appropriateness, which received marginally higher average ratings. The AI-only assessment, in contrast, received slightly higher ratings for Logical Organization and was more consistent with the human raters' scores than the other two choices. The means for the three assessment types (Argumentative Reasoning, Idea Development, and Contextual Appropriateness of Language) did not differ greatly across the three measures. Scores from the hybrid AI-human assessment group fell well within the range of scores from the human-only and AI-only assessment groups, with little variation in outcomes across assessment approaches.

Overall, the descriptive results indicate that all three assessment methods produced fairly similar assessments of students' writing across both cognitive and contextual components. While certain dimensions showed differences, they were relatively minor and therefore appeared to have little practical effect on the descriptive assessment results before inferential analysis.

The second research question examined the level of congruence and agreement between the human-only, AI-only, and hybrid AI-human assessment methods across the six dimensions of writing. ICC was initially analyzed for agreement, and then weighted kappa coefficients were applied to evaluate the agreement between pairs.

Table 3. Intraclass Correlation Coefficients (ICC) among Human-only, AI-only, and Hybrid AI-Human Assessment

Writing Dimension	ICC	95% CI	Interpretation
Logical Organization	.82	.76-.87	Good
Argumentative Reasoning	.74	.66-.81	Moderate
Idea Development	.76	.69-.82	Good
Contextual Appropriateness of Language	.79	.72-.84	Good
Audience Awareness	.71	.63-.79	Moderate
Register and Tone Appropriateness	.73	.65-.80	Moderate

Table 3 below shows intraclass correlation coefficients (ICC) for the six dimensions of writing across the three assessment approaches, ranging from .71 to .82 and suggesting moderate to good agreement among the three assessment approaches. Logical Organization (ICC = .82), Contextual Appropriateness of Language (ICC = .79), and Idea Development (ICC = .76) had the highest level of agreement. Lower ICCs were found for Audience Awareness (ICC = .71), Register and Tone Appropriateness (ICC = .73), and Argumentative Reasoning (ICC = .74).

The results indicate that the three assessment methods tended on average to assess students' writing in a similar way, but that there was not always a strong convergence in the way the various aspects of writing were assessed. In particular, the ICC scores for Audience Awareness and Register and Tone Appropriateness were still somewhat

lower than those for other dimensions involving more contextual interpretation and audience sensitivity, reflecting greater intersubjective variation in dimensions that placed more emphasis on audience interpretation or appropriateness. To further explore how consistent the assessment approaches were, weighted kappa coefficients were computed for each dimension of writing. The results of the study are tabulated in Table 4.

Table 4. Weighted Kappa Coefficients for Pairwise Agreement among Assessment Approaches

Writing Dimension	Human-AI	Human-Hybrid	AI-Hybrid
Logical Organization	.74	.82	.80
Argumentative Reasoning	.61	.77	.73
Idea Development	.66	.79	.76
Contextual Appropriateness of Language	.70	.81	.78
Audience Awareness	.58	.75	.71
Register and Tone Appropriateness	.60	.76	.73

The pattern was generally similar to that found in the ICC analysis, as reflected in the weighted kappa coefficients. Agreement between the human-only and AI-only assessments ranged from fair to moderate (κ = .58 to .74) across the six writing dimensions. The highest agreement between Human and AI was for Logical Organization (κ = .74), while the lowest was for Audience Awareness (κ = .58), indicating greater disagreement between the two independent evaluation processes for this dimension.

Pairwise agreement between Human and Hybrid was somewhat stronger (0.75-0.82) than between AI and Hybrid (0.71-0.80). Due to the comparatively higher agreement between the human and AI assessments, and since the scores in the hybrid were calculated as an average of the scores generated by each assessment, the agreement compared to the other assessments should be interpreted with due care, as some agreement could relate to the construction of the hybrid scores. This pattern of scores across these domains suggests that the hybrid assessment had a common pattern of scores relative to both types of independent assessment. However, these findings should not be interpreted as equivalence between assessment methods, only as overall consistency.

Overall, the findings suggest that the ICC evaluation approaches provided similar results for the three writing dimensions with weighted kappa analyses showing similar results across each writing domain. In contrast, Audience Awareness and Register and Tone Appropriateness were consistently rated lower, indicating that these dimensions are more difficult to evaluate consistently (contextual and communicative).

The third research question explored if significant differences between human-only, AI-only, and hybrid (AI-human) writing assessment methods were found in each of the six writing dimensions. Repeated Measures Analysis of Variance (ANOVA) for each dimension. Bonferroni adjusted pairwise comparisons were conducted to

determine which specific differences occurred for any significant omnibus effects observed.

Table 5. Repeated Measures ANOVA Comparing the Three Assessment Approaches

Writing Dimension	F	p	Partial η^2p
Logical Organization	4.21	.017	.025
Argumentative Reasoning	1.82	.165	.011
Idea Development	2.11	.124	.013
Contextual Appropriateness of Language	2.64	.074	.016
Audience Awareness	5.37	.006	.031
Register and Tone Appropriateness	3.68	.028	.022

The results in Table 5 indicate that participants' scores for Logical Organization, Audience Awareness and Register, and Tone Appropriateness differed significantly from the first study. We did not find statistically significant differences in participants' scores on Argumentative Reasoning, Idea Development or Contextual Appropriateness of Language when compared with the first study. The corresponding computation of effect sizes was all small (partial η^2p = .011-.031), below the proposed value for medium effects (η^2p = .06), which suggested that the observed differences between the assessment approaches - despite being statistically significant - were not particularly large. The effect sizes for these, however, were consistent and relatively small, with a range of partial η^2p values from .011 to .031, suggesting that the differences in the actual effects of the different approaches to assessing the effects were not large. As such, differences in several comparisons reached statistical significance, but should be carefully regarded for clinical relevance.

Due to the repeated measures research design and N (N = 167), statistically significant results can be obtained even if the magnitude of the difference observed is small. Therefore, the results should be interpreted taking into account both effect size and statistical significance. Overall, the results showed that while there were statistical differences in selected domains, each different approach to writing assessment resulted in essentially similar evaluation on the overall scale of the six dimensions of writing. Pairwise comparisons were done and adjusted with the Bonferroni correction to identify the cause of these differences. Table 6 shows the results.

Table 6. Bonferroni-adjusted Pairwise Comparisons

Writing Dimension	Comparison	Mean Difference	p
Logical Organization	Human vs AI	-0.13	.031
	Human vs Hybrid	-0.07	.418
	AI vs Hybrid	0.06	.502
Audience Awareness	Human vs AI	0.19	.004
	Human vs Hybrid	0.09	.276
	AI vs Hybrid	-0.10	.198
Register and Tone Appropriateness	Human vs AI	0.17	.023
	Human vs Hybrid	0.08	.341
	AI vs Hybrid	-0.09	.286

Statistically significant differences were only found in the Bonferroni-adjusted pairwise comparisons for the Human-AI comparisons. Specifically, AI gave them slightly higher scores with Logical Organization (mean difference = 0.13), while human raters gave them marginally higher scores with Audience Awareness (mean difference = 0.19) and Register and Tone Appropriateness (mean difference = 0.17). Further, no statistically significant differences were found after Bonferroni adjustment between the Human-Human or the Human-Hybrid comparisons and the AI-Hybrid comparison.

Importantly, the magnitude of each absolute mean difference of all statistically significant comparisons was less than 0.20 points (on the assessment scale), representing a minimal amount of variability in assessment scores across the 3 assessment methods. These findings align with the small effect sizes reported in Table 5 and suggest that some statistically significant differences may not be meaningful for assessing EFL writing.

Combining the results in Tables 5 and 6 supports the conclusion that the three assessment approaches showed fairly similar assessment results on the six dimensions of writing. Relatively small differences in the points scored for Logical Organization, Audience Awareness, and Register and Tone Appropriateness have been found and represent statistically significant differences but not significantly different assessment outcomes. In contrast, no statistically significant differences were found for Argumentative Reasoning, Idea Development, and Contextual Appropriateness of Language, further highlighting the consistency of the three assessment methods. Altogether, the results show that, apart from limited differences in aspects that demand more sense or structure, human-written, AI-written and hybrid (AI/human) evaluations were highly similar in their overall evaluations of students' writing.

5. Discussion

This study adopts quantitative and comparative methods to answer the question concerning the use of ChatGPT in the field of assessing EFL argumentative writing relative to human-only and hybrid (human-AI) assessment in six cognitive and contextual dimensions. We compared evaluation results, agreement, and score

differences, although methodological variation was limited because all three evaluations used the same analytic rubric. Although several studies have examined all three types of assessment (AI-only, human-only, and hybrid) in a multidimensional approach, only a handful of studies have tested all three types of assessment in comparison to each other. This study builds on previous work and focuses on comparing these three methods rather than suggesting that one is better than the others. Therefore, the results should be interpreted not only as they fit the context of the assessed text but also the assessment's rubric, the genre of writing, the writers, and the ChatGPT setting utilized in this investigation.

The standardized analytic rubrics that are embedded within the three assessment processes seem to be quite similar and could reduce interrater and inter-scorer variance by providing a more explicit means of scoring, as described by human and AI raters. The overall consistency across the various writing measurement tools may reflect this shared evaluation cheat sheet. In general, human raters gave slightly higher scores for audience awareness and register/tone appropriateness and marginally higher scores for logical organization, while ChatGPT received higher scores for register/tone appropriateness. Scores for argumentative reasoning and idea development were not yet significantly different for the three approaches and there was not yet a significant difference in scores for the contextual appropriateness of language. The differences were slight, but the weighted kappa coefficients show some disagreement for a percentage of individual essays, especially in dimensions that involved more interpretation than a literal rendering of the text. Thus, direct comparisons of mean scores across different assessment approaches are not possible. The findings in this study hold significant promise for areas where AI can help humans deliver more accurate assessments of students' work, such as language education and testing. The results are also similar to other studies on the correlation between the scores of AI-generated student evaluations and human-generated student evaluations under a rubric with clear analytic criteria, which showed moderate-to-high correspondence (Yavuz et al., 2025; Li et al., 2024; Tran et al., 2025; Gjorevski et al., 2025). Nevertheless, they must keep in mind the feedback on the students' writing given by AI tools, as well as the prompts' formulation, students' rhetorical representation construction, and students' rhetorical higher-order features' assessment (Lo et al., 2025; Tahir et al., 2025; Li et al., 2025). The findings of the present research suggest, overall, that the written results generated by ChatGPT are somewhat similar to results generated by experienced human writers following a standardized scoring method for some of the dimensions of writing, but are more different from the results of experienced human writers for a few more dimensions that demand contextual and communicative understanding.

It is more efficient to compare the differences observed in the structure of the writings produced by ChatGPT itself. More specifically, from a Cognitive Models of Writing (CMW) perspective (Flower & Hayes, 1981; Bereiter & Scardamalia, 1987), those instances of logical organization, argumentative reasoning, and idea development can be interpreted as observable manifestations of other cognitive processes that occur during the process of writing: planning, organization, and transformation of knowledge. These dimensions are operationalized through the rubric descriptors and begin to provide textual evidence to guide humans and ChatGPT in the same ways. Audience awareness, contextual appropriateness, and register are viewed as social

constructs in systemic functional linguistics (Halliday, 1985), and evaluating them requires interpreting the audience's communicative orientation and the language choices they make in the communicative process. Hence, these lower levels of agreement compared to other dimensions would reveal their interpretative complexities, not an "inherent limitation of AI-assisted assessment", as Lo et al. (2025), Jovic et al. (2025), and Tahir et al. (2025) have previously stated. Overall, these theories suggest variation between assessment approaches from the assessor's perspective and in the complexity of the construct being assessed.

The findings for the second research question showed moderate to good agreement among the ChatGPT-based, human-only, and hybrid assessment approaches across the six writing dimensions. It is evidenced by both the ICC and the weighted kappa values that the scoring pattern derived from both types of instruments generally resembled each other, with the standardized rubric applied as in this study. These levels of agreement are not intended to imply that there is equivalence between assessment methods. The means of essays' overall scores are comparatively typical, indicating that many essays still had main areas of disagreement in audience awareness and appropriateness of register/tone. The results were consistent with the literature on determining similarities and differences between the ChatGPT generated and students' generated assessments of texts using analytic rubrics with clear instructions, as found to have moderate to high agreement (Taşçı, 2025; Zhang, Erlam & Magalhães, 2025; Yavuz et al., 2025), as well as previous findings that suggest that the types of analytic rubrics, complexity of the construct and the prompts, essay characteristics, and degree of agreement may vary (Ataseven et al., 2025; Tahir et al., 2025). From the standpoint of "Technology-Enhanced Language Assessment (Shermis & Burstein, 2013)," and taking into account the results set above, it seems that AI could be utilized to streamline the assessment of measures that are readily interpretable by objective criteria, while human intervention would be needed to interpret measures that require some degree of contextual and possibly rhetorical understanding.

The results for the third research question offer some further insights into the relationship between the three assessment methods. Statistically significant differences were found in logical organization, audience awareness, register, and tone appropriateness, while no differences were found in argumentative reasoning, idea development, or contextual appropriateness of language. However, these effect sizes were small, indicating that, given the conditions under which these tests were administered in the current study, the differences were modest. These are examples showing the significance of distinguishing between significant (statistical means) and significant (practical means). Although the differences in mean marks were not significant or large, individual essays were sometimes marked differently across approaches, as shown by the agreement analysis. Ennadir and Habiballah (2025), Ataseven et al. (2025), and Wolf & Suhan (2025) also proposed that while there are differences in a set of average scores, it may be possible to agree on the scores that are given to individual writing samples in the testing. Therefore, results from AI-based writing assessments should be interpreted as an aggregate of data and the degree of agreement with the individual.

Careful consideration of the hybrid AI - human assessment is also needed. The hybrid scores were in the middle of the range in comparison to the two extremes - AI-only and human-only - and yielded a relatively high level of agreement with both extremes.

However, this is not a case where hybrid assessment is always superior, as the hybrid scores were "composed" of both AI-generated and human-generated assessments. It is important to consider this part of the agreement between the two scores: the score obtained from the hybrid assessment, which in the present instance was relatively high, was interpreted in the light of the procedure used in determining the hybrid scores: some degree of agreement was to be expected, given the hybrid process. However, the results suggest that using AI-generated and human assessments simultaneously is a feasible approach to integrating non-correlated sources of evidence in writing assessment. The findings also align with previous studies suggesting that AI should support, rather than supplant, human evaluators (Ding et al., 2025; Khojasteh et al., 2025; Zhou, 2025). The results therefore don't contradict the use of a hybrid assessment model in other cases; rather, they support it compared with other models (here, the equal-weight hybrid model has been investigated, rather than expressing the superiority of hybrid assessment in general). Meanwhile, the relative value assigned to each source of evidence and how assessments are implemented in the classroom are likely to affect the effectiveness of hybrid assessment, as will the relative importance of AI and human judgments.

The results extend the body of research comparing AI-supported writing assessment with human-performed writing assessment because they include a hybrid method appraised in terms of both cognitive and contextual aspects of writing. Rather, the results indicate that there could be more differences between the approaches depending on the type of writing construct being evaluated and how the writing is being evaluated. The theoretical approaches employed in this study, then, do not explain in more detail why specific dimensions of writing would show greater consistency across approaches; they offer alternative lenses for explaining construct representation.

The results also have practical implications, but these can be addressed in the current study. The results do not endorse replacing human assessment or encourage immediate adoption of AI-supported assessment tasks in institutional settings; rather, they indicate that ChatGPT can serve as an additional tool in specific assessment scenarios, provided certain assessment conditions are met. The benefit AI assessments can offer that may complement the human evaluator is that they can aid with the concealment dimension of writing, which the human evaluator is vital for, since it might assist in outlining parts using a semantic scale where points could be awarded earlier than the unit teacher works out, which leads to the part being scored with analytic elements. When considering the use of AI for assessment, it is important to engage with the technology mindfully, including the issues of construct validity, transparency and equity, as well as the potential impact of the disparity between how well the AI assesses and how well the human assesses when making and remaking.

Lastly, the limitations of the present research will guide where future studies may go next. The patterns observed should be further examined to determine whether the same pattern occurs in other teaching situations, with different learner types, classes, writing styles, levels and writing abilities and LLM types. However, the results should be considered in the context of using one very large language model with a static evaluation question. The behaviors of a large language model may change from version to version of the model and from one prompt design to another. The output of the present study can only be generalized if the specific setting of AI tools used in the

present study is matched. In the future, it would be beneficial to investigate different approaches to modeling hybridization and their effectiveness, examine individual essays with substantial disagreement between AI-generated and human-generated assessments, and explore differences in assessments produced by various LLM models and prompts. This analysis would help clarify when AI for assessment adds value in L2 assessment and improve future construct validation.

Conclusion

AI is beginning to shape language evaluation and could significantly impact how writing is evaluated, but it raises serious questions about validity, transparency, and the need for human judgment. Against this background, this study compared AI-only, human-only, and hybrid (AI + human) assessment across cognitive and contextual dimensions in EFL writing. This research builds on existing studies and knowledge by showing how the interpretations of AI-assisted assessment vary with the writing constructs being assessed, with the assessment procedures used, and with the educational context in which the assessing decisions are made. Together, these findings offer a more balanced view of how AI can be used in writing assessment while providing further construct-based evidence for using AI to supplement assessment.

It is important to keep in mind that the findings in this study are constrained by the methodological context, namely the EFL context, argumentative writing task, rubric, the group of participants, and the method of implementing ChatGPT. Accordingly, there is still a need to validate it in other ways, including other writing contexts, writing genres, learner types and AI systems. Future research should also explore reproducibility of AI-assisted assessment across different versions of large language models and prompting types to strengthen the generalisability and long-term viability of AI-supported writing evaluation. Learning outcomes will be shaped not only by technological evolution but also by evolving assessment frameworks that embody quality criteria such as validity, fairness, transparency and meaningful human oversight in an AI-powered learning environment. This partnership will continue to flourish, is critical to its remaining a beneficial risk to good assessment practice, and will be crucial for AI to not endanger good practice.

Acknowledgement
The author extends their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2026/02/45289).

Bibliographic references
Alam, S. (2026). AI-powered automated speaking scoring and human evaluation: a comparative study of EFL learners' oral proficiency. Arab World English Journal, 17(3), 246-273. https://dx.doi.org/10.24093/awej/vol17no3.14
Alam, S., Usama, M., Shams, A., & Tausif, M. R. (2026). From fear to fluency: a classroom intervention using positive emotional language to reduce foreign language anxiety. Journal of Education Culture and Society, 17(2), 243-269. https://doi.org/10.15503/jecs2026.2.243.269
Asadi, M., Ebadi, S., & Mohammadi, L. (2025). The impact of integrating ChatGPT with teachers' feedback on EFL writing skills. Thinking Skills and Creativity, 57, Article 101766. https://doi.org/10.1016/j.tsc.2025.101766

Ataseven, H., Bökeoğlu, Ö. Ç., & Taşdemir, F. (2025). Can AI assess writing skills like a human? A reliability analysis. Kuramsal Eğitim Bilim Dergisi, 18(4), 757-775. https://doi.org/10.30831/akukeg.1718511
Bereiter, C., & Scardamalia, M. (1987). Knowledge telling and knowledge transforming in written composition. In S. Rosenberg (Ed.), Advances in applied psycholinguistics (Vol. 2, pp. 142-175). Cambridge University Press.
Bouziane, K., & Bouziane, A. (2024). AI versus human effectiveness in essay evaluation. Discover Education, 3(1). https://doi.org/10.1007/s44217-024-00320-6
Chapelle, C. A., & Voss, E. (2016). Twenty years of technology and language assessment. Language Learning & Technology, 20(2), 116-128.
Cheng, H. F., & Dörnyei, Z. (2007). The use of motivational strategies in language instruction: The case of EFL teaching in Taiwan. Innovation in Language Learning and Teaching, 1(1), 153-174. https://doi.org/10.2167/illt048.0
Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.
Cohen, L., Manion, L., & Morrison, K. (2018). Research methods in education (8th ed.). Routledge. https://doi.org/10.4324/9781315456539
Council of Europe. (2001). Common European framework of reference for languages: Learning, teaching, assessment. Cambridge University Press.
Creswell, J. W., & Creswell, J. D. (2018). Research design: Qualitative, quantitative, and mixed methods approaches (5th ed.). SAGE.
Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. Assessing Writing, 18(1), 7-24. https://doi.org/10.1016/j.asw.2012.10.002
Ding, L., Zou, D. & Kohnke, L. (2025). ChatGPT as an automated writing evaluation tool: how students perceive it and how it affects their writing. Education and Information Technologies, 30(2025), 26777-26799. https://doi.org/10.1007/s10639-025-13775-3
Dörnyei, Z. (2007). Research methods in applied linguistics. Oxford University Press.
Du, J., & Nordin, N. R. B. M. (2025). A systematic review of automated writing evaluation (AWE) systems on university students' English writing performance. Forum for Linguistic Studies, 7(11), 615-632. https://doi.org/10.30564/fls.v7i11.11764
Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. Language Assessment Quarterly, 9(3), 270-292.
Ennadir, H., & Habiballah, S. (2025). A comparative study of human and AI-assisted assessment using ChatGPT: The case of Moroccan EFL learners. Arab World English Journal, 16(3), 175-187.
Fathi, J., & Rahimi, M. (2026). Utilising artificial intelligence-enhanced writing mediation to develop academic writing skills in EFL learners: a qualitative study. Computer Assisted Language Learning, 39(1-2), 263-302. https://doi.org/10.1080/09588221.2024.2374772
Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39(2), 175-191. https://doi.org/10.3758/BF03193146
Field, A. (2018). Discovering statistics using IBM SPSS Statistics (5th ed.). SAGE.
Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. College Composition and Communication, 32(4), 365-387.
Gjorevski, A., Li, M. and Cox, T.L. (2025), Exploring the Potential of ChatGPT for Evaluating English Essays in a Criterion-Based Assessment. TESOL Quarterly, 59(2025), 251-279. https://doi.org/10.1002/tesq.70011
Halliday, M. A. K. (1985). An introduction to functional grammar. Edward Arnold.

Hartwell, K., & Aull, L. (2023). Editorial introduction: AI, corpora, and future directions for writing assessment. Assessing Writing, 57(2023), 100769. https://doi.org/10.1016/j.asw.2023.100769
Jovic, M., Papakonstantinidis, S. & Kirkpatrick, R. From red ink to algorithms: investigating the use of large language models in academic writing feedback. Lang Test Asia, 15(59) (2025). https://doi.org/10.1186/s40468-025-00389-2
Khojasteh, L., Soori, A., & Javed, F. (2025). Comparing teacher E-feedback, AI feedback, and hybrid feedback in enhancing EFL writing skills. Technology in Language Teaching & Learning, 7(3), 102626. https://doi.org/10.29140/tltl.v7n3.102626
Knoch, U. (2009). Diagnostic writing assessment: The development and validation of a rating scale (Vol. 17). Peter Lang.
Kobylarek, A. (2026). Key Competences after AI Revolution. Journal of Education Culture and Society, 17(2), 9-15. https://doi.org/10.15503/jecs2026.2.9.15
Kobylarek, A., & Birova, J. (2026). AI as the philosophy of a matrix society. educational implications. Journal of Education Culture and Society, 17(1), 9-18. https://doi.org/10.15503/jecs2026.1.9.18
Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012
Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics, 33(2), 363-374. https://doi.org/10.2307/2529786
Li, J., Huang, J., Wu, W., & Whipple, P. B. (2024). Evaluating the role of ChatGPT in enhancing EFL writing assessments in classroom settings: A preliminary investigation. Humanities & Social Sciences Communications, 11(1). https://doi.org/10.1057/s41599-024-03755-2
Lo, N., Chan, S., & Wong, A. (2025). Evaluating teacher, AI, and hybrid feedback in English language learning: Impact on student motivation, quality, and performance in Hong Kong. SAGE Open, 15(3). https://doi.org/10.1177/21582440251352907
Lu, Q., Yao, Y., Xiao, L., Yuan, M., Wang, J., & Zhu, X. (2024). Can ChatGPT effectively complement teacher assessment of undergraduate students' academic writing? Assessment & Evaluation in Higher Education, 49(5), 616-633. https://doi.org/10.1080/02602938.2024.2301722
Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? Language Testing, 19(3), 246-276.
Pill, J., & Harding, L. (2013). Defining the language assessment literacy gap: Evidence from a parliamentary inquiry. Language Testing, 30(3), 381-402. https://doi.org/10.1177/0265532213480337
Prior, P. (2006). A sociocultural theory of writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), Handbook of writing research (pp. 54-66). Guilford Press.
Shermis, M. D., & Burstein, J. (Eds.). (2013). Handbook of automated essay evaluation: Current applications and new directions. Routledge.
Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86(2), 420-428. https://doi.org/10.1037/0033-2909.86.2.420
Tabachnick, B. G., & Fidell, L. S. (2019). Using multivariate statistics (7th ed.). Pearson.
Tahir, M., Jahrir, A. S., Asrifan, A., & Ariatna, A. (2025). Examining the role of AI-powered writing assistants in enhancing critical thinking in EFL academic writing. Journal of Languages and Language Teaching, 13(4), 1817-1831. https://doi.org/10.33394/jollt.v13i4.15077

Taşçı, S. (2025). Human and AI Scoring of EFL Writing: The Influence of Rubrics and Genre on Reliability. Eğitim Ve Yeni Yaklaşımlar Dergisi, 8(2), 191-210. https://doi.org/10.52974/jena.1785369
Tran, Y. V., Le, T. H. G., Tran, Y. V., & Le, T. H. G. (2025). Evaluating ChatGPT's reliability in grading writing assignments on the EOP learning platform. ICTE Conference Proceedings, 7, 1-19. https://doi.org/10.54855/979-8-9870112-8-7_1
Weigle, S. C. (2002). Assessing writing. Cambridge University Press.
Wolf, M. K., & Suhan, M. (2025). Language assessment and learning through AI technology: An exploratory study on using GPT for young EFL learners' writing. Language Teaching Research Quarterly, 50, 77-100.
Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. British Journal of Educational Technology, 56, 150-166. https://doi.org/10.1111/bjet.13494
Zhou, Y. (2025). Exploring the teacher-student-AI triad in college EFL teaching: A perspective of HITL theory. English Language Teaching, 18(11), 33-43. https://doi.org/10.5539/elt.v18n11p33
Zhang, T., Erlam, R., & de Magalhães, M. B. (2025). Exploring the dual impact of AI in post-entry language assessment: Potentials and pitfalls. Annual Review of Applied Linguistics, 45, 274-293.

Words: 12140
Characters: 85 385 (47 standard pages)

Sohaib Alam, Dr. Ph.D.
Department of English Language and Literature,
College of Sciences and Humanities, Alkharj
Prince Sattam Bin Abdulaziz University
Abdullah Bin Amer Street, 16278,
Kingdom of Saudi Arabia
s.alam@psau.edu.sa
ORCID 0000-0002-9972-9357

Appendix A

Analytic Scoring Rubric for EFL Writing Assessment

The analytic scoring rubric was developed to evaluate EFL argumentative writing across six dimensions comprising three cognitive dimensions (logical organization, argumentative reasoning, and idea development) and three contextual dimensions (contextual appropriateness of language, audience awareness, and register and tone appropriateness). Each dimension was evaluated independently using a five-point scale ranging from **1 (Very Poor)** to **5 (Excellent)**.

Dimension	Score	Descriptor
Logical Organization	**5**	Ideas are sequenced coherently and purposefully throughout the essay. Paragraphs are logically connected, transitions are effective, and the argument progresses smoothly from introduction to conclusion.
	4	Organization is generally logical with only minor lapses in sequencing or transitions. Overall progression remains clear and coherent.
	3	Basic organizational structure is evident, although inconsistencies in sequencing, paragraph development, or transitions occasionally reduce coherence.
	2	Organization is weak. Ideas are frequently disordered, and relationships between paragraphs or ideas are unclear.
	1	No clear organizational structure is evident. Ideas appear random and lack logical progression.
Argumentative Reasoning	**5**	A clear and well-supported position is maintained throughout. Claims are logically justified with convincing evidence, and counterarguments are effectively acknowledged or refuted.
	4	The position is clear and generally supported with logical reasoning. Counterarguments are addressed but not fully developed.
	3	A position is identifiable, but reasoning is inconsistent or insufficiently supported. Counterarguments receive limited attention.
	2	The position is unclear or weakly supported. Reasoning is incomplete, repetitive, or illogical.
	1	No coherent argumentative position is presented, and

Dimension	Score	Descriptor
		reasoning is absent or largely irrelevant.
Idea Development	**5**	Ideas are thoroughly developed with specific explanations, relevant examples, and sufficient elaboration that fully address the writing task.
	4	Most ideas are adequately developed with relevant supporting details, although some points require further elaboration.
	3	Ideas are presented but only partially developed. Supporting examples are limited or insufficiently explained.
	2	Ideas receive minimal development and are supported by few or irrelevant examples.
	1	Ideas are undeveloped, incomplete, or unrelated to the writing task.
Contextual Appropriateness of Language	**5**	Language choices consistently reflect the academic context. Vocabulary, grammatical structures, and expressions are appropriate for formal academic writing.
	4	Language is generally appropriate with only occasional lapses in academic style or word choice.
	3	Language is partially appropriate but demonstrates inconsistent awareness of academic conventions.
	2	Language frequently reflects inappropriate vocabulary, expressions, or grammatical choices for the academic context.
	1	Language is consistently inappropriate for academic writing and substantially limits effective communication.
Audience Awareness	**5**	Writing consistently demonstrates awareness of the intended academic audience through appropriate explanations, tone, and level of detail.
	4	Audience awareness is generally maintained, although occasional lapses occur in adapting information to reader expectations.
	3	Some awareness of audience is evident, but adaptation is

Dimension	Score	Descriptor
		inconsistent or superficial.
	2	Limited consideration is given to the intended audience, resulting in unclear or inappropriate communication.
	1	No evidence of audience awareness is demonstrated throughout the essay.
Register and Tone Appropriateness	**5**	A consistently formal academic register and objective tone are maintained throughout the essay.
	4	Register and tone are generally appropriate with only minor inconsistencies.
	3	Register fluctuates between formal and informal styles, reducing overall consistency.
	2	Register and tone are frequently inappropriate for academic writing.
	1	Register and tone are consistently inappropriate and fail to meet the expectations of academic discourse.

Appendix B

Standardized ChatGPT Evaluation Prompt

The following standardized prompt was used to evaluate each argumentative essay. The same prompt was applied to all essays without modification to ensure consistency across AI-generated evaluations.

Role

You are an experienced EFL writing assessor with expertise in academic writing assessment. Evaluate the following argumentative essay using the analytic scoring rubric provided below.

Assessment Criteria

Evaluate the essay on the following six dimensions.

1. Logical Organization

Assess the overall organization of the essay, including the logical sequencing of ideas, paragraph structure, coherence, and use of transitions.

2. Argumentative Reasoning

Assess the clarity of the writer's position, quality of reasoning, use of supporting evidence, treatment of counterarguments, and logical consistency of the argument.

3. Idea Development

Assess the extent to which ideas are sufficiently explained, elaborated, and supported with relevant examples or details.

4. Contextual Appropriateness of Language

Assess whether vocabulary, grammar, and language choices are appropriate for formal academic writing and consistent with the communicative purpose of the essay.

5. Audience Awareness

Assess the extent to which the writing demonstrates awareness of the intended academic audience through appropriate explanations, level of detail, and communicative effectiveness.

6. Register and Tone Appropriateness

Assess whether the essay consistently maintains an appropriate formal academic register and objective tone.

Scoring Scale

Assign **one score (1-5)** for each dimension according to the following scale:

- **5 = Excellent**
- **4 = Good**
- **3 = Satisfactory**
- **2 = Limited**
- **1 = Poor**

Instructions

1. Read the entire essay before assigning scores.
2. Evaluate each dimension independently.
3. Assign only one whole-number score (1-5) for each dimension.
4. Provide a brief justification (one to three sentences) for each score.
5. Base your evaluation solely on the content of the essay.
6. Do not infer information that is not explicitly present in the text.
7. Do not revise, rewrite, or edit the student's essay.
8. Maintain the same scoring standard throughout all evaluations.

Output Format

Dimension	Score (1-5)	Brief Justification
Logical Organization		
Argumentative Reasoning		
Idea Development		
Contextual Appropriateness of Language		
Audience Awareness		
Register and Tone Appropriateness		

Overall Comments

Provide a concise summary (approximately 80-100 words) highlighting the principal strengths and weaknesses of the essay without altering the assigned scores.