ChatGPT as an Automated Writing Assessment Tool: Inter-rater Alignment and Feedback Agreement

Wilma Villacís Villacís [1], Félix Fernández-Peña [2,*], and Dennis Chicaiza Castillo [2]

[1] Language Center, Universidad Técnica de Ambato, Ambato, Ecuador
[2] Software Engineering, Facultad de Ingeniería en Sistemas, Electrónica e Industrial, Universidad Técnica de Ambato, Ambato, Ecuador
Email: wilmagvillacisv@uta.edu.ec (W.V.V.); fo.fernandez@uta.edu.ec (F.F.-P.); dv.chicaiza@uta.edu.ec (D.C.C.)
*Corresponding author
Manuscript received May 21, 2026; revised June 30, 2026; accepted August 3, 2026; published November 24, 2026

***Abstract*—This mixed-methods study evaluated ChatGPT (GPT-5.3) as an automated writing assessment tool by examining its alignment with human raters in scoring A2-level email writing and the quality of its feedback. Sixty student emails were independently assessed by four human raters and two ChatGPT scoring instances using a Common European Framework of Reference for Languages (CEFR)-aligned analytic rubric. Agreement between ChatGPT and human raters was analyzed using Quadratic Weighted Kappa. Separately, three English as a Foreign Language (EFL) experts qualitatively evaluated the generated feedback. The results show that ChatGPT-human agreement (κ = 0.414) was comparable to human-human agreement (κ = 0.394), while agreement between ChatGPT scoring instances was substantial (κ = 0.661). ChatGPT consistently assigned higher scores and showed closer alignment with lenient evaluators. Expert evaluation indicated partial agreement with the generated feedback, with the main discrepancies concerning organization and language. These findings suggest that ChatGPT provides scoring within the range of human variability but that its feedback still requires human oversight, supporting its use as a complementary writing assessment tool.**

***Keywords*—English as a Foreign Language (EFL) writing, ChatGPT, automated writing assessment, assessment alignment, feedback alignment**

I. Introduction

The advent of Generative Artificial Intelligence (GenAI) has introduced new possibilities for learning and assessing foreign language skills [1, 2]. This technology has provided educators and learners unprecedented capabilities to generate images, producing written content and engaging in human-like interaction [3]. As a result, GenAI has prompted shifts in instructional methodologies across educational contexts [4]. Within language education, these developments have encouraged teachers to adopt GenAI-based tools to optimize teaching and learning processes.

In the field of English as a Foreign Language (EFL) teaching, technological innovation has influenced assessment practice, particularly when it comes to evaluating written production. Writing assessment at the A2 level presents specific challenges, as learners at this stage frequently experience difficulties with accuracy, coherence, and vocabulary range [5].

Historically, writing assessment has been conducted primarily by instructors, drawing on their expert judgment of linguistic features, discourse organization, and sociocultural aspects of language use [4]. To support learners' progress, professors are required to provide frequent and detailed feedback. This process, however, is time-consuming and cognitively demanding [6]. For instance, instructors must evaluate a voluminous amount of student texts, a task that requires sustained concentration and evaluative consistency across ratings [7].

These constraints have drawn scholarly attention to Automated Writing Evaluation (AWE) and Automated Essay Scoring (AES) as alternatives to human evaluation [8]. Effective GenAI-based scoring may mitigate, and potentially eliminate, sources of human variability such as rater fatigue and attentional decline within the assessment process [9].

The integration of AI-based tools in English as a Foreign Language assessment is not entirely new. Applications such as Grammarly, Quillbot, and ProWritingAid have been widely used as automated writing evaluation tools to identify linguistic errors and suggest improvements in learners' texts [10]. Building on these earlier systems, the emergence of Large Language Models (LLMs) has introduced more advanced and unified possibilities for automated feedback and scoring of written production [11].

Among LLM-based applications, ChatGPT has attracted particular attention due to its capacity to perform grammatical and syntactic analysis and to deliver feedback aligned with scoring rubrics. Moreover, it can generate complete written texts in response to prompts [12], and it could assist teachers by offering instant feedback and preliminary scoring based on set criteria, allowing educators to focus more on personalized instruction [13, 14].

Although previous studies have focused on the reliability of AI for essay assessment [8, 15–17], these findings cannot be directly generalized to A2 email writing.

Genre theory views writing as a socially situated practice in which each genre is shaped by its communicative purpose and audience expectations rather than by general writing ability alone [18]. Consequently, essays typically emphasize the development and logical organization of ideas, argumentation, and coherence, whereas emails prioritize interactional goals such as requesting, informing, apologizing, or responding appropriately within authentic communicative contexts. Likewise, models of communicative competence argue that effective language performance extends beyond grammatical accuracy to include discourse, sociolinguistic, and pragmatic competence [19].

From a pedagogical perspective, these distinctions are reflected in the Common European Framework of Reference for Languages (CEFR), which defines A2 writing primarily in terms of functional written interaction. At this level, learners are expected to produce short personal messages and routine correspondence to accomplish everyday communicative purposes rather than develop extended

academic arguments [20]. Therefore, findings obtained from automated assessment of essays cannot be assumed to generalize directly to email writing, as the two tasks differ in both the constructs they assess and the communicative competences they are intended to measure.

Further research is therefore required to determine whether ChatGPT can reliably evaluate email writing at the A2 level accuracy and consistency in comparison with human raters. Accordingly, investigating how ChatGPT assesses A2 learners' writing in relation to human raters using standardized rubrics may yield valuable insights into its pedagogical value, limitations, and ethical considerations. This study addresses the following research question:

RQ1: To what extent do ChatGPT's scores of A2 learners' email writing align with human raters' evaluations?

RQ2: To what extent do EFL experts agree with the feedback generated by ChatGPT on A2 learners' email writing?

II. Literature Review

A. ChatGPT Scoring Reliability

Assessment systems using GenAI are increasingly positioned as solutions to longstanding challenges in Second Language (L2) writing assessment [21–23]. Nevertheless, García-Varela et al. found that when no rubric was supplied, ChatGPT generated fluctuating evaluation criteria and inconsistent grades across sessions [24]. Even under rubric-guided conditions, the model introduced additional scoring dimensions not specified by instructors, raising concerns about construct validity. These findings suggest that LLM-based scoring remains highly sensitive to prompt design and rubric clarity, necessitating rigorous protocol standardization to achieve acceptable reliability.

Meanwhile, Uyar and Büyükahıska examined ChatGPT 4o mini as an AES tool using IELTS Task 2 criteria. While they reported moderate correlations between GenAI and human ratings, they found statistically significant score discrepancies for descriptive and narrative genres [16]. While stronger alignment emerged for analytically structured essays, overall consistency was deemed insufficient for autonomous deployment.

Additionally, Yang [14] examined ChatGPT 3.5's reliability in scoring 82 EFL essays. Although ChatGPT's mean scores did not significantly differ from those of human raters, intra-rater and inter-rater reliability were only moderate, and performance was particularly weak in assessing organization. The two previous studies concluded that ChatGPT should not be considered a fully reliable autonomous scoring tool [15, 16].

Conversely, Bucol and Sangkawong [25] concluded that ChatGPT demonstrated the capacity to conduct evaluations effectively using customized rubrics, achieving levels of consistency comparable to human raters. However, the authors suggested that teachers remain attentive to minor discrepancies, such as difficulties processing complex linguistic features and possible formulaic evaluative patterns.

Similarly, Aydın *et al.* [17] analyzed 590 Turkish L2 essays scored by GPT-4o using a CEFR-aligned rubric. The authors reported strong agreement with human raters (QWK = 0.72; r = 0.73), despite systematic differences such as stricter penalization of morphological and lexical errors [26]. Consistent with these findings, Mizumoto *et al.* [27] reported that ChatGPT aligns more closely with human judgments of linguistic accuracy than conventional AWE tools such as Grammarly. Such findings highlight the potential of GenAI to meet human evaluative standards in L2 writing assessment.

B. Instructional Impact of ChatGPT Feedback

Shi *et al.* [23], in an eleven-week mixed-methods study conducted with intermediate language proficiency levels, found that students receiving ChatGPT feedback achieved significantly higher post-test writing scores compared to those using traditional AWE systems. Nonetheless, this ChatGPT group also reported lower perceptions of L2 writing ability. The authors associated this trend with increased tool dependence and reduced beliefs in creativity and agency. This finding aligns with Siekmann, Parr, and Busse [28], who emphasized the role of feedback in supporting deeper writing competencies, particularly structure and coherence. Likewise, Mahapatra [29] demonstrated that guided use of ChatGPT for self and peer-assessment led to greater and more sustained writing improvements than traditional instruction alone. It was underscored that ChatGPT has a direct positive impact on student learning since it can provide feedback which is tailored to each student. Nonetheless, the authors suggested investigating the impact of ChatGPT in an extended intervention period.

Ding and Zou [30] conducted a study with the participation of 77 graduate students. They reported generally positive student perceptions of the use of ChatGPT in writing practice, particularly regarding surface-level issues with grammar and vocabulary feedback. However, ChatGPT′s ability to address the more complex aspects of writing was reported as limited. The study also emphasized the importance of prompt training.

Similarly, Gozali *et al.* [10] found that combining ChatGPT with rule-based tools such as Grammarly and Quillbot fostered feedback literacy through complementary functions. These include (1) supporting idea development, (2) meaning-level processing, and (3) rule-based tools addressing grammatical form. Nevertheless, concerns about over paraphrasing, feedback reliability, and ethical use remained. Alolaywi [26] also examined how QuillBot assesses writing in comparison to human graders. The findings indicated that the bot effectively provides consistent and highly accurate detection of mechanical writing features. It also highlights that QuillBot can serve as an efficient tool to support teachers and enable them to focus more on higher-order writing skills.

Previous research has also examined the impact of GenAI-supported feedback on higher-order thinking. Hao and Razali [31] considered that integrating AWE (Pigai) and ChatGPT feedback over 16 weeks resulted in significantly greater gains in analysis, evaluation, and creation compared to AWE alone. While GenAI feedback encouraged deeper cognitive engagement, the authors emphasize the importance of guided human oversight. Similarly, Zhou [15] found that ChatGPT provides comprehensive discourse-level feedback in IELTS writing, specifically in coherence and

argumentation, whereas traditional AWE tools are perceived as more effective for lower-order concerns such as grammar and mechanics. Their study concludes that ChatGPT shows pedagogical promise when integrated within human guided feedback frameworks.

From a broader perspective, Zhang [9] situates ChatGPT within the evolution of Automated Written Corrective Feedback (AWCF), noting its ability to move beyond the form-focused limitations of earlier systems. Nevertheless, their review warns that without principled pedagogical integration, such affordances may foster overreliance and reduce engagement with higher-order composing processes.

Li [32] synthesized theoretical and empirical evidence to identify four principal instructional applications of ChatGPT in L2 writing: (1) GenAI-assisted idea generation, (2) teacher-mediated feedback support, (3) collaborative learner processing of GenAI output, and (4) GenAI-supported assessment. The study emphasized that effectiveness depends on explicit scaffolding and careful teacher mediation to mitigate risks such as misinformation, construct underrepresentation, and academic integrity concerns.

Evidence from related educational domains further underscores the complementary roles of GenAI and human expertise. Deriba *et al*. [33] found no short-term differences between GenAI-generated and instructor feedback in teaching inclusive design. However, instructor feedback led to significantly stronger long-term retention. Likewise, a systematic review of 42 empirical studies in Ref. [34] concluded that while LLMs perform well on closed-ended and short-answer tasks, they struggle with complex, open-ended, and subjective assignments. Their findings support hybrid human supervised models, emphasizing that LLM effectiveness depends on task type, rubric specificity, language context, and model version.

Empirical evidence on ChatGPT's writing assessment remains heterogeneous. Most studies have focused on scoring performance and reliability, while less attention has been given to the agreement between expert evaluators and ChatGPT-generated feedback, particularly in A2-level email writing. This represents a notable gap in literature, as the pedagogical validity of GenAI-generated feedback depends on its alignment with expert judgment, not only on its ability to produce scores.

III. Materials and Methods

This study was based on an experiment to evaluate ChatGPT powered by the GPT-5.3 model as an automated writing assessment tool. This version was selected because it was the latest model available at the time of the study. The experiment comprised two phases. In the first phase, addressing RQ1, sixty students from an A2-level English course (39 women and 21 men aged 19 to 25) were asked to write a 120-word informal email. Their writings were rated by each of four university professors, denoted as reviewers H_1, H_2, H_3, and H_4. Three of the professors were non-native speakers and one was a native speaker of English. Each email was also assessed twice by ChatGPT, denoted as AI instances AI_1 and AI_2. Using a multi-layer AI approach allowed us to examine intra-model consistency. Crucially, the objective was not to estimate the full distribution of the model's responses but rather determine whether repeated assessments yielded comparable results. Consequently, two independent evaluations per email were considered adequate for the purposes of this study.

Each reviewer utilized a Delphi validated rubric. The Delphi panel consisted of three experts in teaching English pedagogy. The use of the Delphi method is consistent with recent research adopting the same approach [35]. Nine items were evaluated using a Likert scale, which corresponded to the following categories: inter-rater reliability, global clarity, relevance, CEFR alignment, score differentiation, the adequacy of descriptors for content, communicative achievement, organization, and language. Experts all agree that the rubric was accurately designed.

Each request to ChatGPT was based on this prompt: "Act as an expert English teacher with more than 30 years of experience and assess these writings from A2 learners with the rubric. This is the task: Write a 120 word- text about this topic. TOPIC: You had a terrible shopping experience. Write an email to your friend. In your email, you should: "Say who you were with." "Describe what happened." "Say how you felt." "Say what people should do when shopping." "I send you the rubric".

The scores given by all raters were recorded in an Excel database. Inter-rater agreement was analyzed using weighted linear Kappa coefficients. To examine differences in scoring severity across evaluators, the Friedman test was applied, followed by post hoc Wilcoxon signed-rank tests with Holm-Bonferroni adjustment.

In the second phase, responding to RQ2, three experts who had a combined twenty years of EFL teaching experience, provided a qualitative assessment of ChatGPT feedback. Each expert had attained advanced degrees in language pedagogy and carried direct experience working with A2 level learners.

Each EFL teaching expert reviewed three A2-level student emails and the corresponding ChatGPT-generated feedback. They were asked to assess if the ChatGPT-generated feedback was accurate, aligned with the established rubric, and provided actionable constructive feedback for the learner.

Expert judgments were classified into predefined agreement categories: full agreement, partial agreement, and disagreement, and provided their reasons in an interview. The data was analyzed using a convergent design. Descriptive statistics were used to quantify agreement levels, while qualitative analysis of expert explanations identified patterns of convergence and divergence in the evaluation of ChatGPT-generated feedback.

IV. Result and Discussion

A. Inter-rater Agreement in Overall Scoring

The inter-rater agreement was measured in terms of relations among humans H_x–H_y (x, y ∈ [1,4]), AI instances AI_i–AI_j (i, j ∈ [1,2]), and among H_x–AI_i. In the overall evaluation, as shown in Table 1, consistency between the two ChatGPT scoring instances was substantial (κ = 0.661), indicating that the system maintained a relatively stable pattern when applying the same rubric to the same 60 e-mail texts. Agreement between the AI_i and H_x was heterogeneous: the closest alignment was with H_2 (κ = 0.561 for AI_1 and

κ = 0.540 for AI_2) and with H_4 (κ = 0.463 for AI_1 and κ = 0.555 for AI_2), while H_1 was the most distant evaluator compared to both ChatGPT scoring instances (κ = 0.235 and κ = 0.229). Among humans, the highest overall agreement was between H_2 and H_3 (κ = 0.583), confirming that variability is not limited to the H_x–AI_i contrast, but is also part of the inter-human-behavior itself.

Table 1. Weighted linear Kappa for overall score by evaluator pair

Pair	K(Kappa)	Interpretation
AI_1–AI_2	0.661	Substantial
AI_1–H_1	0.235	Fair
AI_1–H_2	0.561	Moderate
AI_1–H_3	0.391	Fair
AI_1–H_4	0.463	Moderate
AI_2–H_1	0.229	Fair
AI_2–H_2	0.54	Moderate
AI_2–H_3	0.342	Fair
AI_2–H_4	0.555	Moderate
H_1–H_2	0.411	Moderate
H_1–H_3	0.361	Fair
H_1–H_4	0.261	Fair
H_2–H_3	0.583	Moderate
H_2–H_4	0.457	Moderate
H_3–H_4	0.293	Fair

Note. Descriptive interpretation of K: slight (<0.21), fair (0.21–0.40), moderate (0.41–0.60, and substantial (0.61–0.80).

B. Comparison between AI-human Discrepancies and Human-human Discrepancies

To assess whether differences within the ChatGPT scoring exceeded human variability, two indicators were synthesized: mean Kappa by type of relationship and average absolute difference in total score. Table 2 shows that the pattern is clear: in the overall score, mean AI_i–H_x agreement (κ = 0.414) is very close to mean H_x–H_y agreement (κ = 0.394). Similarly, the average absolute difference in total score is practically identical between AI_i–H_x pairs (1.013 points) and H_x–H_y pairs (1.022 points). This equivalence supports the idea that discrepancies between ChatGPT and humans are, in general magnitude, similar to those observed within the human group itself.

Table 2. Comparative synthesis of agreement by type of relationship

Dimension	Type of relationship	Mean Kappa	Min.	Max.	N. of pairs
Overall score	AI_i–H_x	0.414	0.229	0.561	8
Overall score	H_x–H_y	0.394	0.261	0.583	6
Overall score	AI_i–AI_j	0.661	0.661	0.661	1
Content	AI_i–H_x	0.351	0.139	0.521	8
Content	H_x–H_y	0.34	0.171	0.534	6
Content	AI_i–AI_j	0.721	0.721	0.721	1
Communicative achievement	AI_i–H_x	0.4	0.186	0.561	8
Communicative achievement	H_x–H_y	0.38	0.266	0.627	6
Communicative achievement	AI_i–AI_j	0.658	0.658	0.658	1
Organization	AI_i–H_x	0.384	0.165	0.585	8
Organization	H_x–H_y	0.398	0.277	0.551	6
Organization	AI_i–AI_j	0.596	0.596	0.596	1
Language	AI_i–H_x	0.43	0.282	0.576	8
Language	H_x–H_y	0.343	0.252	0.428	6
Language	AI_i–AI_j	0.576	0.576	0.576	1

In the overall score, the AI_i–H_x and H_x–H_y pairs show very similar absolute discrepancies as presented in Table 3, which corroborates what was observed in the analysis of the previous table.

Table 3. Comparative indicators for the overall score by type of relationship

Type of relationship	Mean Kappa	Kappa Range	Mean Absolute Difference in Overall score	Range of Absolute Difference	N
H_x–H_y	0.394	0.261–0.583	1.022	0.467–1.883	6
AI_i–H_x	0.415	0.229–0.561	1.013	0.067–1.950	8
AI_i–AI_j	0.661	0.661–0.661	0.383	0.383–0.383	1

Note: The mean absolute difference corresponds to the average absolute value of the score differences between the compared pairs.

C. Systematic Differences between Evaluators: Friedman Test

According to Table 4, the Friedman test was significant in the total score and in all four rubric dimensions, showing that the six evaluators did not grade with the same severity. Effect sizes, expressed through Kendall's W, were high in all cases, especially in Content (W = 0.489) and in the total score (W = 0.381). Average ranks show a stable structure: AI_2 and H_4 consistently occupy the pole of higher scores, while H_1 and H_3 tend to occupy the stricter pole. H_2 appears as an intermediate evaluator, though closer to ChatGPT scoring instances in communicative achievement and language.

Table 4. Friedman results for total score and each rubric dimension

Var.	N	χ^2	Df	P	Kendall's W	Mean Rank order
Total	60	114.406	5	<0.001	0.381	AI_2 (4.84) > H_4 (4.60) > AI_1 (3.93) > H_2 (3.05) > H_3 (2.40) > H_1 (2.18)
C	60	146.849	5	<0.001	0.49	AI_2 (4.62) > H_4 (4.53) > AI_1 (4.25) > H_2 (3.00) > H_3 (2.67) > H_1 (1.94)
Cm	60	77.842	5	<0.001	0.259	AI_2 (4.40) > H_4 (3.95) > AI_1 (3.87) > H_2 (3.68) > H_3 (2.85) > H_1 (2.25)
O	60	97.428	5	<0.001	0.325	AI_2 (4.72) > AI_1 (4.15) > H_4 (3.98) > H_2 (2.98) > H_3 (2.76) > H_1 (2.42)
L	60	73.366	5	<0.001	0.245	H_4 (4.66) > AI_2 (4.03) > AI_1 (3.48) > H_2 (3.40) > H_1 (3.04) > H_3 (2.39)

Note: *C*: Content. *Cm*: Communicative achievement. *O*: Organization. *L*: Language. N = 60 email texts.

D. Post Hoc Wilcoxon Comparisons in Total Score

As detailed in Table 5, the post hoc analysis shows that the overall differences detected by the Friedman Test are not distributed homogeneously. In the total score, neither AI_1 nor AI_2 differ significantly from H_4 after Holm–Bonferroni adjustment, while both do differ from H_1, H_2, and H_3. Among teachers, H_1 and H_3 do not differ from each other, allowing identification of a relatively strict pole versus a higher or more lenient pole formed by H_4 and the two ChatGPT scoring instances.

E. Post Hoc Synthesis by Dimension: Convergences and Divergences

When disaggregated by criteria, proximity between ChatGPT scoring instances and humans changes. Table 6

indicates that in Content and Organization, both ChatGPT scoring instances regroup with H_4; in Communicative achievement, AI_1 converges with H_2 and H_4, while AI_2 converges with H_4; in Language, ChatGPT proximity redistributes and no longer concentrates exclusively on H_4. This variation confirms that AI_i–H_j alignment is multidimensional and depends on the rubric criterion.

Table 5. Post hoc Wilcoxon comparisons for total score (Holm–Bonferroni adjustment)

Pair	Z	Adjusted p (Holm)	R	Mean diff. (E1−E2)	Direction
AI_2–H_3	−6.461	<0.001	0.841	1.483	AI_2 > comparator
AI_2–H_1	−6.074	<0.001	0.819	1.95	AI_2 > comparator
AI_2–H_2	−5.643	<0.001	0.775	1.0	AI_2 > comparator
AI_1–H_3	−5.596	<0.001	0.755	1.1	AI_1 > comparator
H_1–H_4	−5.571	<0.001	0.744	−1.883	H_4 > comparator
AI_1–H_1	−5.305	<0.001	0.722	1.567	AI_1 > comparator
H_3–H_4	−4.978	<0.001	0.659	−1.417	H_4 > comparator
H_2–H_4	−4.096	<0.001	0.542	−0.933	H_4 > comparator
H_1–H_2	−4.032	<0.001	0.559	−0.95	H_2 > comparator
AI_1–H_2	−3.469	= 0.003	0.501	0.617	AI_1 > comparator
AI_1–AI_2	−3.467	= 0.003	0.5	−0.383	AI_2 > comparator
H_2–H_3	−3.281	= 0.004	0.455	0.483	H_2 > comparator
H_1–H_3	−2.322	= 0.061	0.319	−0.467	H_3 > comparator
AI_1–H_4	−1.779	= 0.150	0.242	−0.317	H_4 > comparator
AI_2–H_4	−0.616	= 0.538	0.088	0.067	AI_2 > comparator

Note: A negative Z value indicates that the first evaluator in the pair tended to assign higher scores than the second, according to the reported direction.

Table 6. Pairs without significant differences in Wilcoxon post hoc by variable

Variable	AI pairs without difference	Human pairs without difference	Substantive interpretation
Total	AI_1–H_4; AI_2–H_4	H_1–H_3	AI converges primarily with H_4; while H_1 and H_3 represent stricter scoring profiles.
Content	AI_1–AI_2 AI_1–H_4 AI_2–H_4	H_2–H_3	AI_1 and AI_2 cluster with H_4; H_2 and H_3 also show similarity.
Communicative achievement	AI_1–AI_2 AI_1–H_2 AI_1–H_4 AI_2–H_4	H_2–H_4	AI_1 converges with H_2 and H_4; AI_2 converges with H_4; H_2 and H_4 are also similar.
Organization	AI_1–AI_2 AI_1–H_4 AI_2–H_4	H_1–H_2 H_1–H_3 H_2–H_3	AI again clusters with H_4; among humans, H_1, H_2 and H_3 do not differ significantly.
Language	AI_1–AI_2 AI_1–H_1 AI_1–H_2 AI_2–H_2 AI_2–H_4	H_1–H_2 H_1–H_3	AI proximity is redistributed: AI_1 aligns with H_1 and H_2; AI_2 aligns with H_2 and H_4.

Note: *C*: Content. *Cm*: Communicative achievement. *O*: Organization. *L*: Language. Only non-significant pairs after the Holm-Bonferroni adjustment are reported, as these support the discussion of statistical convergence.

F. Agreement of Experts with ChatGPT-generated Feedback

The results presented in Table 7 indicate that agreement between EFL experts and ChatGPT-generated feedback is predominantly partial. While more than half of the evaluations (55.6%) reflect partial agreement, full alignment is rare (11.1%), and one-third of cases (33.3%) show complete disagreement. Qualitative analysis indicates that discrepancies are concentrated in specific rubric dimensions, particularly organization and language. ChatGPT consistently overestimates textual cohesion and applies a more tolerant evaluative criterion for grammatical accuracy, prioritizing interpretability of meaning over clarity.

Table 7. Expert agreement with feedback generated by ChatGPT based on rubric

Email	Expert	Agreement	Main Rubric Dimension(s)	Key Issue
1	1	PA	Content / Organisation / Language	Text partly on task; weak cohesion; grammar affects clarity.
1	2	PA	Language	Feedback useful but too advanced for A2 learners.
1	3	PA	Content / Communicative Achievement / Organisation	AI overestimates scores despite poor cohesion.
2	1	D	Organisation / Language / Communicative Achievement	AI overestimates organization and ignores comprehension problems.
2	2	D	Organisation / Language	AI assigns overly high scores despite clear issues.
2	3	PA	Organisation	Run-on sentences affect coherence (rubric gap).
3	1	PA	Language	Grammar and vocabulary reduce clarity vs AI “meaning inferred”.
3	2	FA	All	Feedback clear, useful and tailored to the student.
3	3	D	Language	AI ignores grammar and applies criteria beyond rubric.

Note: FA: full agreement. PA: partial agreement. D: disagreement. Categories were determined by classifying expert evaluations of ChatGPT-generated feedback.

Additional discrepancies arise from limitations in the rubric, such as the absence of explicit criteria for run-on sentences, and from issues of pedagogical appropriateness, where feedback is not always adapted to the A2 level. These results indicate that although ChatGPT can approximate expert judgment at a general level, its feedback lacks consistent alignment with rubric-based evaluation, especially textual organization and cohesion, and in pedagogical effectiveness of the feedback.

A core source of divergence lies in the evaluative criterion applied to linguistic errors, with ChatGPT adopting a more permissive stance that prioritizes meaning inferability, whereas experts emphasize clarity, organization and communicative effectiveness.

This study examined ChatGPT as an automated writing assessment tool. By analyzing how scores compared to human raters in A2-level email writing, the study indicates that ChatGPT operates within the range of variability that characterizes human writing assessment. These findings are supported by the close similarity between AI_1–H_1 agreement ($\kappa = 0.414$) and human-human agreement ($\kappa = 0.394$), as well as by the nearly identical score discrepancies (1.013 vs. 1.022 points). These results suggest that differences between ChatGPT scoring instances and human raters reflect the inherent subjectivity of writing assessment, rather than a fundamental inconsistency introduced by the model.

Most notably, however, the results also show that ChatGPT does not reproduce human judgment uniformly. Instead, it tends to align with more lenient evaluative profiles, particularly H_4 and, in some dimensions, H_2, whereas H_1 and H_3 represent stricter scoring patterns. These findings indicate that ChatGPT follows a consistent evaluative orientation, characterized by greater leniency, rather than producing randomly variable judgments. A similar pattern is observed in its internal consistency, as the substantial agreement between scoring instances ($\kappa = 0.661$) indicates that the model applies the rubric in a stable and predictable manner.

These results can be explained by the way ChatGPT processes writing features. The model appears to rely more on general textual cues and meaning inferability than on stricter linguistic or discourse-level criteria. This trend becomes evident in the evaluation of organization. ChatGPT tends to overestimate coherence based on surface-level structure, even when human raters identify issues in logical flow and cohesion. Likewise, in the language dimension, the model prioritizes whether the message is understandable, while human raters emphasize clarity, accuracy, and ease of processing. This difference in evaluative focus explains the tendency of ChatGPT to assign higher scores.

Another factor influencing these differences is the rubric itself. As observed in the results, certain aspects of writing quality, such as run-on sentences or discourse features, are not explicitly captured by the descriptors. This may lead both human raters and ChatGPT scoring instances to interpret these elements differently. Therefore, some variability should be understood as a limitation of the assessment instrument, rather than only of the model.

In terms of feedback, the results show predominantly partial agreement with expert evaluations. Although ChatGPT provides detailed responses, its feedback is not always tailored to A2 level learners. In particular, misalignment is evident in aspects related to textual organization and cohesion, where ChatGPT tends to overgeneralize or overlook discourse-level issues. These findings suggest that, although the model can approximate expert judgment at a general level, its feedback does not completely align with rubric-based evaluation and is not instructionally effective.

Overall, these findings indicate that ChatGPT captures key aspects of A2 writing performance, but applies a more lenient evaluative standard, especially in content and organization. These observations align with previous research highlighting that AI-based assessment systems can achieve evaluative consistency but may differ from human raters in interpretive depth. From a practical perspective, these results support the use of ChatGPT as a complementary assessment tool or secondary rater, rather than as a replacement for human evaluation.

In line with these results, the study supports a hybrid approach to writing assessment, where the consistency and efficiency of AI are combined with the interpretive and pedagogical expertise of human raters to ensure valid and meaningful evaluation.

V. Conclusion

This study examined the role of ChatGPT in writing assessment by evaluating its agreement with human raters in assessing A2-level email writing. Although the findings indicate that ChatGPT operated within the variability of human evaluation, the AI model demonstrated leniency by assigning higher scores, particularly in content and organization. Furthermore, ChatGPT prioritized message comprehensibility over strict linguistic and discourse-level accuracy.

Differences are influenced not only by the model's tendency to provide overly positive evaluations and higher scores than human raters, but also by the limitations of the rubric, which does not fully capture certain discourse features. GenAI-generated feedback showed only partial agreement with expert evaluations and was not always pedagogically appropriate for A2 learners, as it did not consistently provide clear and actionable guidance.

From a practical perspective, human expertise remains essential to refine scoring standards, interpret results, and ensure pedagogical relevance. Therefore, these findings suggest that a hybrid assessment model integrating GenAI and human judgment may provide a practical approach to EFL writing assessment when human oversight is maintained. At the same time, these findings do not establish construct validity, as comparable agreement represents only one source of validity evidence and should be complemented by investigations of score meaning and consequential validity.

Future research should include larger and more diverse samples, a wider range of tasks and a greater diversity of proficiency levels. Rubric design, which facilitates ChatGPT-human alignment, also remains an important aspect for further research. In addition to examining the validity of the resulting assessment, further studies should investigate how GenAI-generated feedback and other LLMs can be effectively integrated into classroom practice.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

W.V.V. developed the research idea, F.F.-P wrote the methodology, D.C.C analyzed the data, W.V.V. wrote the manuscript, F.F.-P. supervised and guided the research, D.C.C proofread the manuscript; W.V.V. and F.F.-P did the writing review and editing. All authors approved the final version.

Funding

This research was funded by Dirección of Investigación y Desarrollo, from Universidad Técnica de Ambato, research

project PFISEI37- Metodología para el Uso de la Inteligencia Artificial Generativa en la Educación Superior.

REFERENCES

[1] Y. Wang and Y. Guo, "Investigating the interplay of Chinese English as foreign language learners' growth mindset, Enjoyment, adaptability, and engagement in AI-Mediated informal digital learning of English," *Journal of Computer Assisted Learning*, vol. 42, no. 1, e70189, 2026. doi:10.1002/jcal.70189

[2] D. Wood and S. H. Moss, "Evaluating the impact of students' generative AI use in educational contexts," *Journal of Research in Innovative Teaching & Learning*, vol. 17, no. 2, pp. 152–167, 2024. doi: 10.1108/JRIT-06-2024-0151

[3] W. Charpentier-Jiménez, "Assessing artificial intelligence and professors' calibration in English as a foreign language writing courses at a Costa Rican public university," *Actual. Investig. Educ.*, vol. 24, no. 1, pp. 1–25, 2024. doi: 10.15517/aie.v24i1.55612

[4] A. Dhamija and D. Dhamija, "Understanding teachers' perspectives on ChatGPT-generated assignments in higher education," *Journal of Interdisciplinary Studies in Education*, vol. 14, no. 1, pp. 38–62, 2025. doi: 10.32674/ptf9yd75

[5] C. Uyar and D. Büyükahıska, "Artificial intelligence as an automated essay scoring tool: A focus on ChatGPT," *International Journal of Assessment Tools in Education*, vol. 12, no. 1, pp. 20–32, 2025. doi: 10.21449/ijate.1517994

[6] T. Jansen, L. Höft, L. Bahr, J. Fleckenstein, J. Möller, and J. Meyer, "Comparing generative AI and expert feedback to students' writing: Insights from student teachers," *Psychologie in Erziehung und Unterricht*, no. 1, pp. 80–92, 2024. doi: 10.2378/peu2024.art08d

[7] D. Alonzo, J. M. V. Abril, and C. Z. Oo, "The use of artificial intelligence in English language assessment: Empirical evidence and future directions," *Critical Studies in Teaching and Learning*, vol. 13, no. 2, pp. 108–132, 2025. doi: 10.14426/cristal.v13i2.2723

[8] M. Berijanian, C. Shaltry, and D. Colbry, "Evaluating the evaluators: Metrics for automated essay feedback generation," *International Journal of Information and Education Technology*, vol. 16, no. 4, pp. 894–901, 2026. doi: 10.18178/ijiet.2026.16.4.2561

[9] Y. Zhang, "Incorporating ChatGPT as an automated written corrective feedback tool into L2 writing class," *Journal of Language Teaching*, vol. 4, no. 4, pp. 22–34, 2024. doi: 10.54475/jlt.2024.024

[10] I. Gozali, A. R. T. Wijaya, A. Lie, B. Y. Cahyono, and N. Suryati, "ChatGPT as an Automated Writing Evaluation (AWE) tool: Feedback literacy development and AWE tools' integration framework," *JALT CALL Journal*, vol. 20, no. 1, pp. 1–22, 2024. doi: 10.29140/jaltcall.v20n1.1200

[11] R. Baskara, "Integrating ChatGPT into EFL writing instruction: Benefits and challenges," *International Journal of Education and Learning*, vol. 5, no. 1, pp. 44–55, 2023. doi: 10.31763/ijele.v5i1.858

[12] J. Escalante, A. Pack, and A. Barrett, "AI-generated feedback on writing: Insights into efficacy and ENL student preference," *International Journal of Educational Technology in Higher Education*, vol. 20, no. 1, pp. 1–20, 2023. doi: 10.1186/s41239-023-00425-2

[13] Y. Y. Munaye, W. Admass, Y. Belayneh, M. Molla, and M. Asmare, "ChatGPT in education: A systematic review on opportunities, challenges, and future directions," *Algorithms*, vol. 18, no. 6, 352, 2025. doi: 10.3390/a18060352

[14] Y. Yang, "The reliability of using ChatGPT in rating EFL writings," *Shanlax International Journal of Education*, vol. 12, no. 4, pp. 49–59, 2024. doi: 10.34293/education.v12i4.7855

[15] Y. Zhou, "The effectiveness of automated written corrective feedback on L2 learners' revision outcomes: A case for ChatGPT," *International Journal of New Developments in Education*, vol. 5, no. 25, pp. 58–65, 2023. doi: 10.25236/ijnde.2023.052511

[16] Y. Wang, J. Huang, L. Du, Y. Guo, Y. Liu, and R. Wang, "Evaluating large language models as raters in large-scale writing assessments: A psychometric framework for reliability and validity," *Computers and Education: Artificial Intelligence*, vol. 9, 100481, 2025. doi: 10.1016/j.caeai.2025.100481

[17] B. Aydın, T. Kışla, N. T. Elmas, and O. Bulut, "Automated scoring in the era of artificial intelligence: An empirical study with Turkish essays," *System*, vol. 133, 103784, 2025. doi: 10.1016/j.system.2025.103784

[18] D. Dressen-Hammouda, "From genre theory to indexical framework: Making the invisible visible," *English for Specific Purposes*, vol. 84, pp. 174–180, 2026. doi: 10.1016/j.esp.2026.06.010

[19] E. Nyamekye and W. A. F. Soré, "Exploring the influence of emotion regulation on the social communicative competence of language students in higher education: a PLS-SEM approach," *Discover Psychology*, vol. 6, 22, 2026. doi: 10.1007/s44202-025-00550-1

[20] Council of Europe, *Common European Framework of Reference for Languages: Companion Volume*, Strasbourg, France: Council of Europe Publishing, 2020.

[21] J. Weidlich, F. Gotsch, K. Schudel, C. Marusic-Würscher, J. Mazzarella, H. Bolten, D. Bütler, S. Luger, B. Wohlfender, and K. M. Merki, "Teacher, peer, or AI? Comparing effects of feedback sources in higher education," *Computers and Education Open*, vol. 9, 100300, 2025. doi: 10.1016/j.caeo.2025.100300

[22] P. Y. Liew and I. K. T. Tan, "On automated essay grading using large language models," in *Proc. 2024 8th Int. Conf. Computer Science and Artificial Intelligence (CSAI),* 2025, pp. 204–211. doi: 10.1145/3709026.3709030

[23] H. Shi, C. S. Chai, S. Zhou, and S. Aubrey, "Comparing the effects of ChatGPT and automated writing evaluation on students' writing and ideal L2 writing self," *Computer Assisted Language Learning*, vol. 39, 2025. doi: 10.1080/09588221.2025.2454541

[24] F. García-Varela, M. Nussbaum, M. Mendoza, C. Martínez-Troncoso, and Z. Bekerman, "ChatGPT as a stable and fair tool for automated essay scoring," *Education Sciences*, vol. 15, no. 8, 946, 2025. doi: 10.3390/educsci15080946

[25] J. L. Bucol and N. Sangkawong, "Exploring ChatGPT as a writing assessment tool," *Innovations in Education and Teaching International*, vol. 62, no. 3, pp. 867–882, 2025. doi: 10.1080/14703297.2024.2363901

[26] Y. Alolaywi, "Assessing writing outcomes: A comparative study of artificial intelligence versus traditional assessment techniques," *Arab World English Journal*, no. 3, pp. 118–131, 2026. doi: 10.24093/awej/ai3.8

[27] A. Mizumoto, N. Shintani, M. Sasaki, and M. F. Teng, "Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment," *Research Methods in Applied Linguistics*, vol. 3, no. 2, 100116, 2024. doi: 10.1016/j.rmal.2024.100116

[28] L. Siekmann, J. M. Parr, and V. Busse, "Structure and coherence as challenges in composition: A study of assessing less proficient EFL writers' text quality," *Assessing Writing*, vol. 54, 100672, 2022. doi: 10.1016/j.asw.2022.100672

[29] S. Mahapatra, "Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study," *Smart Learning Environments*, vol. 11, no. 1, 9, 2024. doi: 10.1186/s40561-024-00295-9

[30] L. Ding, D. Zou, and L. Kohnke, "ChatGPT as an automated writing evaluation tool: How students perceive it and how it affects their writing," *Education and Information Technologies*, vol. 30, no. 18, pp. 26777–26799 2025. doi: 10.1007/s10639-025-13775-3

[31] H. Hao, A. B. Razali, and R. Zuo, "Exploring the impact of integrated AWE and generative AI feedback on Chinese EFL undergraduates' higher-order thinking in argumentative writing," *SAGE Open*, vol. 16, no. 1, pp. 1–12, 2026. doi: 10.1177/21582440251413884

[32] M. Li, "Leveraging ChatGPT for second language writing feedback and assessment," *International Journal of Computer-Assisted Language Learning and Teaching*, vol. 14, no. 1, pp. 1–11, 2024. doi: 10.4018/IJCALLT.360382

[33] G. Deriba, M. Saqr, and M. Tukiainen, "Does the impact of GenAI persist in learning? Comparing AI and instructor feedback on accessibility," *Universal Access in the Information Society*, vol. 25, no. 1, 2026. doi: 10.1007/s10209-025-01268-6

[34] M. Jukiewicz and M. Wyrwa, "Can ChatGPT replace the teacher in assessment? A review of research on the use of large language models in grading and providing feedback," *Applied Sciences*, vol. 16, no. 2, 680, 2026. doi: 10.3390/app16020680

[35] P. Villacrés, "Evaluating web interface component adequacy in information management systems: A hybrid performance–heuristic approach," *Journal of Emerging & Sustainable Smart Engineering*, vol. 1, no. 1, pp. 21–47, 2025. doi: 10.31243/jesse.uta.v1i1.2984.2025

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).