Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evaluating Human–AI Rating Agreement in Foundation EFL Writing: Evidence From Analytic Rubric Scoring in Kuwait

Adel Basheer AlAbed

alabed.a@gust.edu.kw

Gulf University for Science & Technology

Article

Keywords: artificial intelligence, EFL writing assessment, human rating, AI rating, analytic rubric, Foundation English, Kuwait

Posted Date: October 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11497492/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Artificial intelligence (AI) is increasingly used to support writing assessment, yet it remains unclear how closely AI-generated scores correspond with human judgment in Foundation English as a foreign language (EFL) contexts. This study examined agreement between AI-generated and human ratings of Foundation EFL writing in Kuwait. The dataset comprised 692 writing assessment records from 162 students enrolled in ENGL097 and ENGL098. Overall scores and six analytic criteria (topic sentence, body and development, conclusion, organization, vocabulary, and grammar) were analyzed using descriptive statistics, Pearson and Spearman correlations, intraclass correlation coefficients (ICCs), mean differences, error indices, and Bland–Altman analysis. AI scores showed very high agreement with the average of two human raters (r = .992; absolute-agreement ICC = .991). The mean difference was 0.24 points on a 50-point scale, the mean absolute error was 0.89 points, and 93.1% of AI ratings fell within two points of the human average. Criterion-level ICCs ranged from .953 to .965, with vocabulary lowest and conclusion highest. A significant proportional bias indicated that disagreement varied across the score range. The findings suggest that AI scoring can provide consistent support for Foundation EFL writing assessment when applied through a clearly defined analytic rubric, although human oversight remains essential.

Introduction

Writing assessment occupies a central position in English language education because writing performance reflects several dimensions of language ability at once, including idea development, organization, vocabulary, and grammatical control. In English as a foreign language (EFL) contexts, where learners are still acquiring linguistic knowledge and academic writing conventions, evaluating writing is demanding, and growing class sizes and more frequent assessment have increased the need for efficient yet dependable scoring. Large language models can interpret written instructions, process analytic rubrics, and generate evaluative judgments, which has raised the question of whether artificial intelligence (AI) can support or partly automate writing assessment (Mizumoto & Eguchi, 2023; Pack et al., 2024; Tate et al., 2024).

Evidence on AI scoring is encouraging but uneven. Agreement varies by model, prompt, task, rubric, and learner population, and many studies emphasize overall score relationships rather than agreement across individual rubric criteria. Most research has also examined standardized essays, placement tests, or relatively advanced university writers, whereas evidence from Foundation-level EFL learners, and from Kuwait and the wider Gulf region in particular, remains limited. Because AI scores are only meaningful when they correspond with trained human judgment under a shared rubric, context-specific, criterion-level evidence is needed before such tools are adopted in Foundation programs.

The present study therefore examines agreement between AI-generated and human ratings of Foundation EFL writing in Kuwait at two levels: overall scores and six analytic criteria. It treats agreement

as more than correlation by combining indices of association, absolute agreement, and score difference, and it uses agreement between two human raters as a benchmark.

Literature Review

From Automated Writing Evaluation to Generative AI Scoring

Early automated writing evaluation systems identified linguistic features, provided corrective feedback, and generated scores through computational models. Research showed that such systems can support second language writing (Cotos, 2011; Li et al., 2015), but their value depends on how they are used and on the quality of the judgments they produce (Chapelle et al., 2015; Ranalli et al., 2017). Reviews and meta-analyses report beneficial but variable effects across tools, contexts, learners, and writing components (Fan & Ma, 2022; Ngo et al., 2024; Nunes et al., 2022), which cautions against assuming that automated tools work uniformly.

Large language models have expanded what is possible. Mizumoto and Eguchi (2023) used a GPT-based model to score 12,100 TOEFL11 essays and found useful reliability and accuracy. Pack et al. (2024) compared GPT-3.5, GPT-4, Claude 2, and PaLM 2 on 119 placement-test essays; GPT-4 performed best, but quality varied across models and rating occasions. Tate et al. (2024) found substantial ChatGPT–human agreement across several corpora of secondary student writing and suggested usefulness for lower-stakes formative purposes, whereas Shermis (2025) found meaningful alignment that was not stable enough for unqualified high-stakes use. Scoring procedures also matter. Kim (2025) showed that rubric detail, scoring rationales, and examples influenced GPT-4 reliability and alignment with human scores, and Huang et al. (2026) similarly found that careful prompting improved GPT-4o accuracy. Many-facet Rasch studies of automated scoring (Chan et al., 2023; Yamashita, 2024) further illustrate the value of treating automated scoring as one rater within a shared measurement framework rather than as an objective standard.

Human Judgment as the Reference and the Meaning of Agreement

Human ratings are the usual reference, but human raters are not perfectly consistent. A rubric alone does not guarantee valid or reliable judgments (Rezaei & Lovorn, 2010), and raters may apply rating scale categories differently across analytic domains (Wind, 2020). Studies that include more than one human rater (Pack et al., 2024; Yavuz et al., 2025) therefore provide a more defensible benchmark than a single teacher's score. Agreement must also be defined carefully: two sources may rank students similarly while assigning systematically different scores, so correlation should be complemented by absolute agreement and by the size and direction of score differences. Kim (2025), for example, found strong internal consistency in GPT-4 but only moderate agreement with human scores, showing that self-consistency and agreement with humans are distinct questions. Positive evidence has been reported for ChatGPT in EFL contexts (Bucol & Sangkawong, 2025; Gjorevski, 2025). Alkhofi (2026) found a strong correlation (r = .816) between calibrated ChatGPT-4o scores and teacher scores of Saudi university

writing, yet ChatGPT was relatively generous with weaker texts whereas teachers awarded relatively higher scores to stronger texts.

Analytic Rubrics and Criterion-Level Agreement

Analytic rubrics divide writing into separate components and are useful for diagnosis and feedback, which is especially relevant in Foundation EFL programs. In Kuwait, students and instructors have viewed rubrics positively because they clarify expectations and support consistent grading (Goodwin & Kirkpatrick, 2023). Criteria differ, however, in how observable they are. Grammar and vocabulary rest on identifiable linguistic evidence, whereas development, coherence, and organization require judgments about relationships among ideas. Consistent with this, Yavuz et al. (2025) found close ChatGPT–human alignment for grammar but variation across rubric domains, and Mizumoto et al. (2024) found strong alignment with human judgments of linguistic accuracy. More interpretive dimensions appear harder: Jin (2025) reported remaining gaps in the evaluation of source use even after calibration, and Zhu and Dong (2025) found significant ChatGPT–instructor differences in content and communicative achievement in Business English writing. Alid and Almegrena (2025) likewise compared AI and human essay evaluation across categories such as content, vocabulary, organization, and accuracy. These studies justify examining agreement criterion by criterion rather than assuming one stable level of accuracy across a rubric.

Systematic Scoring Differences and Model Behavior

AI scores can correlate strongly with human scores while being systematically more lenient or severe. Jin (2025) found that an uncalibrated ChatGPT configuration was the strictest rater and that calibration reduced but did not eliminate differences, showing that severity is not a fixed property of a model. Findings also do not generalize automatically from one system to another, and reliability can change across rating occasions (Pack et al., 2024); an AI scoring system is therefore not a fixed instrument, and the model, prompt, rubric, and procedure must be documented. Strong agreement would show that a system reproduces important aspects of human evaluation under the conditions studied, not that it is valid across all tasks, populations, or model versions.

Foundation EFL Learners and the Kuwait and Gulf Context

Foundation learners often produce short texts with limited vocabulary, unstable grammar, weak cohesion, and unclear paragraph structure, and such writing may challenge human and automated raters alike. Findings from standardized essays or advanced university writers therefore cannot be assumed to transfer. Research in Kuwait shows established rubric use (Goodwin & Kirkpatrick, 2023) and rapid student uptake of AI tools such as ChatGPT, Grammarly, and Google Translate (Alghasab, 2025), but it has focused mainly on writing practices rather than score correspondence. Work in Saudi Arabia has examined ChatGPT feedback (Alsofyani & Barzanji, 2025) and AI–human scoring (Alid & Almegrena, 2025; Alkhofı, 2026), yet evidence on Foundation EFL students in Kuwait and across locally meaningful analytic criteria remains scarce.

The Present Study

The study is organized around parallel AI and human assessment of the same writing samples under a shared analytic rubric. Agreement is examined first for total scores and then for six criteria, and it is conceptualized as both association and numerical consistency, with human–human agreement as a benchmark. Two research questions guided the study.

RQ1. To what extent do AI-generated writing scores agree with human raters' scores in Foundation EFL writing assessment?

RQ2. To what extent does AI–human rating agreement vary across the six rubric criteria: topic sentence, body and development, conclusion, organization, vocabulary, and grammar?

Method

Design

This study used a quantitative comparative agreement design. Its purpose was to determine the consistency of different rating sources scoring the same performances rather than to test an intervention. Each writing sample was scored by an AI system and by two human raters using the same analytic rubric, so the text was constant while the rating source varied. Because correlation indicates association rather than agreement, complementary indicators were used (Chan et al., 2023; Pack et al., 2024; Yamashita, 2024). Human interrater agreement served as a benchmark because human raters also vary in severity and in their interpretation of rubric categories (Rezaei & Lovorn, 2010; Wind, 2020).

Context and Participants

The study was conducted in the Foundation English program at Gulf University for Science and Technology (GUST) in Kuwait. Data came from two courses: ENGL097 (approximately A2–B1) and ENGL098 (approximately B1–lower B2). The dataset comprised 692 writing assessment records from 162 students collected during the 2024, 2025, and 2026 academic years. Of these records, 318 were from ENGL097 and 374 from ENGL098, distributed across 12 course sections. The 162 students (83 male, 51.2%; 79 female, 48.8%) were all Kuwaiti and aged 17 to 19 years. Students could contribute several records (227 student–course records), so the number of observations exceeded the number of participants, and the individual record was the unit of analysis. The records comprised 346 midterm writing exam paper and 346 final exam papers from both ENGL097 and ENGL098.

Writing Tasks and Rubrics

Samples were drawn from official midterm and final examinations. Students wrote one paragraph of approximately 150 to 200 words in one hour in either of two genres, opinion or narrative. Examinations

were completed on paper; some handwritten responses were scored by the AI system directly, and others were manually transcribed without correcting grammar, spelling, punctuation, vocabulary, organization, or content. Records were included if the sample was produced in ENGL097 or ENGL098 in 2024–2026 as part of an official examination, contained sufficient text for scoring, and had both AI and human ratings. Records were excluded if they were incomplete, illegible, missing ratings, duplicated, or otherwise unsuitable for comparison.

Genre-specific analytic rubrics were used for opinion and narrative writing, each assessing the same six dimensions: topic sentence (clarity of the controlling idea), body and development (relevance, elaboration, and support), conclusion (effectiveness of the concluding sentence), organization (logical sequencing and coherence), vocabulary (appropriateness and range), and grammar (accuracy). Criterion scores were summed to a 50-point total score, which was used in the agreement analyses.

AI Scoring Procedure

The AI assessment system was implemented through a customized ChatGPT based grading bot developed for Foundation EFL writing assessment. The bot incorporated the analytic writing rubrics and standardized scoring instructions used in the study. A standardized prompt containing the relevant genre-specific rubric instructed the system to assign a score for each of the six criteria and a total score without rewriting or correcting the student's text. Each sample was scored twice. The repeated scoring procedure was used to examine the consistency of the AI-generated ratings and reduce reliance on a single scoring event. Because model version, prompting, and rubric presentation influence AI scores (Kim, 2025; Pack et al., 2024), the full rubrics are provided in the Appendix.

Human Raters

Two experienced EFL instructors with backgrounds in Foundation-level teaching and writing assessment scored the samples using the same genre-appropriate rubrics. They reviewed and discussed the rubrics before scoring, then scored independently and were blind to the AI scores. Their scores were retained separately to allow interrater analysis, and the mean of the two was used as the primary human reference.

Data Analysis

Records were anonymized and screened for duplicates, missing values, transcription inconsistencies, and out-of-range scores. Analyses were conducted in SPSS Version 30. Descriptive statistics were computed for AI scores, each human rater, and the human average. Human interrater agreement was examined with Pearson correlation and absolute-agreement intraclass correlation coefficients (ICCs). For RQ1, AI scores were compared with each human rater and with the human average using Pearson and Spearman correlations, absolute-agreement ICCs, mean difference, mean absolute error (MAE), and root mean square error (RMSE). Paired-samples t tests and Cohen's dz assessed differences between

rating sources. Bland–Altman analysis estimated mean bias and 95% limits of agreement, and the correlation between pair means and difference scores was examined as an indicator of proportional bias. For RQ2, Pearson correlation, absolute-agreement ICC, mean difference, MAE, and paired comparisons were computed for each criterion, with human interrater ICCs as a benchmark. A supplementary analysis compared agreement across courses. Significance was evaluated at $\alpha = .05$, but effect sizes and the practical size of differences were emphasized because the large number of observations could make very small differences statistically significant. Because some students contributed multiple records, observations were not independent, and inferential statistics should be interpreted with this in mind.

Results

Descriptive Statistics and Human Rater Agreement

As shown in Table 1, mean total scores were similar across rating sources. The two human raters showed very high agreement ($r = .967$; ICC = .967), with a mean difference of −0.17 points, MAE of 1.63, and RMSE of 2.07. The difference was statistically significant, $t(691) = -2.11$, $p = .036$, but small.

Table 1 Descriptive Statistics for AI and Human Total Scores

Rating source	M	SD
AI rating	38.96	8.32
Human Rater 1	38.64	7.99
Human Rater 2	38.81	8.02
Average human rating	38.73	7.94

Note. N = 692 records. Scores are out of 50.

Overall AI–Human Agreement (RQ1)

AI scores agreed very strongly with both human raters and with their average (Table 2). Agreement was highest with the human average ($r = .992$, $\rho = .986$, ICC = .991), with a mean difference of 0.24 points, MAE of 0.89, and RMSE of 1.12. Paired-samples tests showed significant but small differences between AI and Human Rater 1, $t(691) = 5.15$, $p < .001$, $dz = 0.20$; Human Rater 2, $t(691) = 2.96$, $p = .003$, $dz = 0.11$; and the human average, $t(691) = 5.69$, $p < .001$, $dz = 0.22$. In practical terms, AI scores fell within 0.5 points of the human average for 39.6% of records, within 1 point for 64.7%, within 2 points for 93.1%, within 3 points for 99.6%, and within 5 points for 100%.

Table 2 Agreement Between AI and Human Total Scores

Comparison	r	ρ	ICC	Mean diff.	MAE	RMSE	dz
Human 1 vs. Human 2	.967	—	.967	−0.17	1.63	2.07	—
AI vs. Human 1	.981	.969	.979	0.32	1.31	1.66	0.20
AI vs. Human 2	.987	.979	.986	0.15	1.10	1.37	0.11
AI vs. average human	.992	.986	.991	0.24	0.89	1.12	0.22

Note. N = 692. ρ = Spearman correlation; ICC = absolute-agreement intraclass correlation coefficient; mean diff. = AI minus human (Human 1 minus Human 2 in the first row); MAE = mean absolute error; RMSE = root mean square error. Dashes indicate statistics not reported.

Bland–Altman Analysis

The mean bias was 0.24 points, indicating that AI scores were slightly higher than the human average. The 95% limits of agreement ranged from −1.91 to 2.38 points. Differences did not depart significantly from normality (Shapiro–Wilk $p = .286$). However, the average score was positively related to the AI–human difference ($r = .346$, $p < .001$), indicating some proportional bias: the size of disagreement varied across the score range.

Criterion-Level Agreement (RQ2)

Agreement was very high for all six criteria (Table 3). ICCs ranged from .953 for vocabulary to .965 for conclusion, followed by body and development (ICC = .962), so the range was narrow. Mean differences were very small (0.00 to 0.10 points), and MAE ranged from 0.25 to 0.46. Paired comparisons were significant for topic sentence ($p = .026$), body and development ($p < .001$), organization ($p = .040$), and vocabulary ($p < .001$), but not for conclusion ($p = .976$) or grammar ($p = .195$). The largest difference, 0.10 points for body and development, concerned a criterion worth 12 points, so these significant results appear to reflect the large number of observations rather than substantial disagreement. Human interrater ICCs were lower and also varied across criteria, from .819 for vocabulary to .864 for body and development, with MAE between 0.50 and 0.86.

Table 3 Criterion-Level Agreement Between AI and Average Human Ratings

Criterion	AI M	Human M	r	ICC	Mean diff.	MAE	p	Human ICC
Topic sentence	5.57	5.54	.959	.959	0.03	0.26	.026	.852
Body and development	9.21	9.10	.964	.962	0.10	0.46	< .001	.864
Conclusion	5.47	5.48	.965	.965	0.00	0.25	.976	.855
Organization	6.31	6.28	.959	.958	0.03	0.30	.040	.844
Vocabulary	6.23	6.18	.954	.953	0.05	0.33	< .001	.819
Grammar	6.18	6.16	.959	.959	0.02	0.30	.195	.830

Note. N = 692. Human M is the mean of the two human raters. Mean diff. = AI minus human average; p = paired-samples test of the mean difference; Human ICC = absolute-agreement ICC between the two human raters. ICC = intraclass correlation coefficient; MAE = mean absolute error.

Agreement Across Course Levels

Agreement was similarly high in both courses (Table 4).

Table 4 AI and Human Agreement by Course Level

Course	n	AI M	Human M	r	ICC	MAE
ENGL097	318	39.04	38.85	.991	.990	0.82
ENGL098	374	38.90	38.63	.993	.991	0.95

Note. Human M is the mean of the two human raters. ICC = absolute-agreement intraclass correlation coefficient; MAE = mean absolute error.

Discussion

The findings show very high agreement between AI-generated and human ratings of Foundation EFL writing. This is consistent with studies reporting that large language models can produce scores that correspond closely with human ratings under structured conditions (Gjorevski, 2025; Mizumoto & Eguchi, 2023; Pack et al., 2024; Tate et al., 2024). Agreement here was higher than the moderate relationships reported by Kim (2025) and the correlation reported by Alkhofi (2026), which may reflect differences in task, rubric, model, and scoring procedure. The use of ICCs, error indices, and Bland–Altman analysis shows that correspondence extended beyond rank order to the scores themselves.

Human interrater agreement provided an important benchmark. AI agreement with each human rater (ICC = .979 and .986) was at least as high as agreement between the two human raters (ICC = .967), and

criterion-level human ICCs (.819–.864) were lower than AI–human ICCs (.953–.965). Even with a shared rubric, human raters therefore showed non-trivial variation, consistent with Rezaei and Lovorn (2010) and Wind (2020), and AI–human differences should be read against this variation. Agreement with the human average is also expected to exceed agreement with a single rater because averaging reduces rater-specific variation.

Criterion-level agreement was uniformly high. Differences among criteria were too small to support strong claims about which dimensions are easier for AI, although vocabulary showed the lowest agreement for both AI–human and human–human comparisons. These results contrast with studies reporting domain variation (Mizumoto et al., 2024; Yavuz et al., 2025) and difficulty with interpretive dimensions (Jin, 2025; Zhu & Dong, 2025). The structured, single-paragraph tasks and rubrics used here may make even development and organization more observable, a possibility that future research should test. Small but significant mean differences for four criteria were practically negligible, so statistical significance should be interpreted alongside effect size. The proportional bias also indicates that strong overall agreement does not mean AI and human ratings behave identically at every performance level (Alkhofi, 2026). Agreement was similar in ENGL097 and ENGL098, suggesting that it was not confined to one proficiency level within this program.

Pedagogical Implications

AI scoring may serve as a supplementary tool for preliminary scoring, formative assessment, second marking, and rapid feedback, and criterion-level scores may help instructors identify specific areas of difficulty. However, AI scores should support rather than replace professional judgment, particularly when decisions affect progression or final course outcomes (Shermis, 2025; Tate et al., 2024). The results also underscore the importance of clear analytic rubrics, which help both human and AI raters score consistently (Goodwin & Kirkpatrick, 2023).

Limitations and Future Research

Several limitations should be considered. The data came from one Foundation program at one university, which limits generalizability. Students contributed multiple records, so observations were not independent and significance tests should be interpreted cautiously. Only opinion and narrative paragraphs were examined. Some handwritten responses were scored directly and others were transcribed, which may have affected AI processing. Finally, the findings apply to the specific model, prompt, and rubric used, because AI performance may vary with model version, prompt design, rubric wording, and system updates (Kim, 2025; Pack et al., 2024). Future studies should replicate the analysis across institutions, proficiency levels, genres, and AI models and versions, examine score-level bias and handwritten versus typed input, and investigate whether AI-supported assessment improves feedback quality, scoring efficiency, and writing development over time.

Conclusion

This study found very high agreement between AI-generated and human ratings of Foundation EFL writing in Kuwait, both overall and across six analytic criteria. AI scoring can correspond closely with human judgment when applied through a structured analytic rubric. Evidence of proportional bias and small criterion-level differences, however, indicates that AI should not be treated as a fully independent replacement for human assessment. In Foundation EFL writing, AI is best positioned as a supportive tool that complements instructor judgment, improves efficiency, and provides an additional source of scoring evidence.

Declarations

This study was reviewed and approved by the Institutional Review Board (IRB) of Gulf University for Science and Technology (GUST), Kuwait. All procedures involving human participants were conducted in accordance with the ethical standards of the institutional review board and relevant ethical guidelines.

Author Contribution

I, A.B.A, conceived and designed the study, developed and configured the AI writing assessment system, collected and prepared the data, conducted the analysis, interpreted the results, and wrote and revised the manuscript.

References

1. Alghasab, M. B. (2025). English as a foreign language secondary school students' use of artificial intelligence tools for developing writing skills: Unveiling practices and perceptions. Cogent Education, 12(1), Article 2505304.
2. Alid, J. K., & Almegrena, R. M. (2025). Evaluating the quality of AI feedback: A comparative study of AI and human essay grading. Innovations in Education and Teaching International, 62(6), 1858–1873.
3. Alkhofi, A. (2026). Can ChatGPT score ESL writing? A correlation analysis between teacher and GenAI scores. Language Teaching Research. Advance online publication.
4. Alsofyani, A. H., & Barzanji, A. M. (2025). The effects of ChatGPT-generated feedback on Saudi EFL learners' writing skills and perception at the tertiary level: A mixed-methods study. Journal of Educational Computing Research, 63(2), 431–463.
5. Bucol, J. L., & Sangkawong, N. (2025). Exploring ChatGPT as a writing assessment tool. Innovations in Education and Teaching International, 62(3), 867–882.
6. Chan, K. K. Y., Bond, T., & Yan, Z. (2023). Application of an automated essay scoring engine to English writing assessment using many-facet Rasch measurement. Language Testing, 40(1), 61–85.
7. Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. Language Testing, 32(3), 385–405.

8. Cotos, E. (2011). Potential of automated writing evaluation feedback. CALICO Journal, 28(2), 420–459.
9. Fan, N., & Ma, Y. (2022). The effects of automated writing evaluation feedback on students' English writing quality: A systematic literature review. Language Teaching Research Quarterly, 28, 53–73.
10. Gjorevski, N. (2025). Exploring the potential of ChatGPT for evaluating English essays in a criterion-based assessment. TESOL Quarterly, 59(S1), S251–S279.
11. Goodwin, R., & Kirkpatrick, R. (2023). Using rubrics to improve writing skills: A study in Kuwait. Language Testing in Asia, 13, Article 17.
12. Huang, Y., Palermo, C., & Wilson, J. (2026). Accuracy and fairness of generative AI in automated essay scoring: Comparing GPT-4o, feature-based models, and human raters. Assessing Writing, 69, Article 101047.
13. Jin, H. (2025). When AI meets source use: Exploring ChatGPT's potential in L2 summary writing assessment. System, 133, Article 103737.
14. Kim, Y. (2025). Automated essay scoring with GPT-4 for a local placement test: Investigating prompting strategies, intrarater reliability, and alignment with human scores. TESOL Quarterly, 59(S1), S318–S329.
15. Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation feedback in ESL writing instruction. Journal of Second Language Writing, 27, 1–18.
16. Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. Research Methods in Applied Linguistics, 2(2), Article 100050.
17. Mizumoto, A., Shintani, N., Sasaki, M., & Teng, M. F. (2024). Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment. Research Methods in Applied Linguistics, 3(2), Article 100116.
18. Ngo, T. T. N., Chen, H. H. J., & Lai, K. K. W. (2024). The effectiveness of automated writing evaluation in EFL and ESL writing: A three-level meta-analysis. Interactive Learning Environments, 32(2), 727–744.
19. Nunes, A., Cordeiro, C., Limpo, T., & Castro, S. L. (2022). Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020. Journal of Computer Assisted Learning, 38, 599–620.
20. Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. Computers and Education: Artificial Intelligence, 6, Article 100234.
21. Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. Educational Psychology, 37(1), 8–25.
22. Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. Assessing Writing, 15(1), 18–39.

23. Shermis, M. D. (2025). Using ChatGPT to score essays and short-form constructed responses. Assessing Writing, 66, Article 100988.
24. Tate, T. P., Steiss, J., Bailey, D., Graham, S., Moon, Y., Ritchie, D., Tseng, W., & Warschauer, M. (2024). Can AI provide useful holistic essay scoring? Computers and Education: Artificial Intelligence, 7, Article 100255.
25. Wind, S. A. (2020). Do raters use rating scale categories consistently across analytic rubric domains in writing assessment? Assessing Writing, 43, Article 100416.
26. Yamashita, T. (2024). An application of many-facet Rasch measurement to evaluate automated essay scoring: A case of ChatGPT-4.0. Research Methods in Applied Linguistics, 3(3), Article 100133.
27. 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(1), 150–166.
28. Zhu, Y., & Dong, M. (2025). Decoding the raters' mind: A metadiscourse investigation of ChatGPT versus EFL instructors in Business English writing assessment. Studies in Educational Evaluation, 87, Article 101541.

Figures

098 Narrative Paragraph Rubric Name:_____ID_____ Course&Section:_____Date:_____**T1**_____ **T2**_____

Underline ONLY the descriptors that apply

	Topic sentence 7	**Body** 12	**Concluding sentence** 7	**Organization & logic** 8	**Vocabulary & Spelling** 8	**Grammar & punctuation** 8
Exceeds Expectations	**7** TS is very well-written; introduces a strong main idea; limits or focuses the topic Topic very well paraphrased; (NOT copied from the prompt)	**12** Body has very well-developed, sufficient and relevant beginning, middle and end details; very well-developed, sufficient and relevant examples & explanations which strongly support the topic sentence	**7** -The concluding sentence (why the story is important/significant) is very well-written, paraphrased, logical, convincing, and well developed	**8** Ideas flow smoothly & A wide range of accurate, less common connectors and transitions; correct formatting and paragraphing	**8** Wide range of academic vocab/ collocations with high-vocab & Conveys precise meaning in spite of minor errors; mostly accurate word choice, form, and colloc; rare spelling errors that do not affect meaning	-A range of sentence types - mostly accurate and meaning clear & -Rare structure, punctuation or capitalization errors that do not affect meaning **8**
Good to Very good Meets Expectations	**5.25** TS is well-written; introduces the main idea and limits or focuses the topic; Topic minimally paraphrased (partially copied from the prompt)	**9.5** Body has some well-developed, sufficient and relevant beginning, middle and end details; Body has some well-developed examples and explanations	**5.25** -The concluding sentence (why the story is important/significant) is well-written but only partially paraphrased;	**6** Logical organization & A range of connectors and transitions; an error may cause only temporary confusion OR Minor formatting or paragraphing errors	Adequate range of vocab/ collocations for the topic & includes some high vocab but needs more; a few errors of word choice, form, or colloc but meaning clear; occasional spelling errors but meaning clear **6**	**6** -Simple structures very well controlled -Complex structures attempted but very few faulty OR -Some structure, punctuation or capitalization errors but meaning clear
Needs Improvement	**3** TS barely mentions the main idea /topic; does not limit or focus the topic, is not easy to understand, or needs minor modification -Topic fully copied from the prompt (no attempt to paraphrase);	**6** Body has poorly-written, insufficient, undeveloped beginning, middle or end details; a few ideas are confusing, unrelated, OR Body needs minor development	**3.5** -The concluding sentence (why the story is important/significant) is fully copied from TS, lacks precision, or needs minor modification;	**4** Weak organization of ideas; ideas don't always flow OR limited, repetitive, very basic, or a few faulty connectors and transitions OR a few cause confusion OR faulty formatting or paragraphing	**4** Basic vocab or too much repetition; frequent errors of word choice, form, or colloc; frequent spelling errors; meaning may be confusing or unclear	**4** Limited or mostly simple structures; major problems in structures; frequent sentence structure, punctuation or capitalization errors; meaning sometimes confusing or unclear
Unsatisfactory	**2** TS is unclear, faulty or irrelevant, or with no specific focus; OR Main idea (topic) not mentioned (missing) -Zero if TS missing	**3** Body is missing details or has unrelated or many unclear ideas; Or unclear how details relate to the main idea of the narrative -Zero if less than 3 lines/sentences	**2** -The concluding sentence (why the story is important/significant) is unclear, vague, or unrelated to the main idea of the paragraph -**Zero** if missing	**2** Very poor organization of ideas; connectors and transitions rare, repetitive, faulty, or missing; not enough to evaluate	Sentences are essentially translated from Arabic; very limited control of word choice, form, or colloc; errors cause some strain for the reader; dominated by spelling errors; meaning is mostly unclear; not enough to evaluate **2**	Very limited control of sentence structure; dominated by punctuation or capitalization errors; errors may cause some strain for the reader; not enough to evaluate **2**

0 if absent, paragraph plagiarized or not submitted **Score:** /50 **Average** /50 **Total** /100 **% (Pass ≥ 70%)**

ENGL 098 Paragraph Rubric / November 20, 2023

Figure 1

Writing Rubric – Narrative

098 W1 Opinion Paragraph Rubric Name:_____ ID_____ Course&Section _____ Date _____ T1_____ T2_____

	Topic sentence 7	Body 12	Concluding sentence 7	Organization & logic 8	Vocabulary & Spelling 8	Grammar & punctuation 8
Exceeds Expectations	**7** TS is very well-written; introduces a strong opinion with a clear controlling idea; topic very well paraphrased - not copied from the prompt	**12** 3 reasons; very well labelled, well-developed, and relevant; all details are very well-developed, and relevant; examples & explanations strongly support the reasons and the topic sentence	**7** Concluding sentence / opinion very well paraphrased; summary of 3 reasons accurate and very well paraphrased; relevant and logical	Ideas flow smoothly **8** & A wide range of accurate, less common connectors and transitions; correct formatting and paragraphing	**8** Wide range of academic vocab/ collocations with high-vocab & Conveys precise meaning in spite of minor errors; mostly accurate word choice, form, and colloc; rare spelling errors that do not affect meaning	-A range of sentence **8** types - mostly accurate and meaning clear & -Rare structure, punctuation or capitalization errors that do not affect meaning
Good to Very good Meets Expectations	**5.25** TS well-written; introduces a clear opinion with a clear controlling idea; topic minimally paraphrased (partially copied from the prompt);	**9.5** 3 reasons; clearly labelled, easy to understand; minor details clearly support reasons (explanations, examples, etc.); ideas sufficiently developed; one reason or detail needs better explanation or minor modification	**5.25** Concluding sentence / opinion and topic partially paraphrased (not fully copied from TS); Summary attempts to paraphrase the 3 reasons adequately, in one sentence, and in the correct order	**6** Logical organization & A range of connectors and transitions; an error may cause only temporary confusion OR Minor formatting or paragraphing errors	Adequate range of vocab/ **6** collocations for the topic & includes some high vocab but needs more; a few errors of word choice, form, or colloc but meaning clear; occasional spelling errors but meaning clear	**6** -Simple structures very well controlled -Complex structures attempted but very few faulty OR -Some structure, punctuation or capitalization errors but meaning clear
Needs Improvement	**3** TS / opinion needs minor modification; or Topic fully copied from prompt (no attempt to paraphrase); Controlling idea missing, unclear, or needs minor modification	**6** Reasons: Poorly labelled, unclear, less than 3; or one irrelevant. Details: 1-2 details are missing, poorly-written, unclear, or unrelated to reasons	**3.5** Concluding sentence / opinion and topic fully copied from TS, is missing opinion, or needs minor modification; 3 reasons fully copied from body; or summary copied fully poor, unclear, too long, or needs minor modification	**4** Weak organization of ideas; ideas don't always flow OR limited, repetitive, very basic, or a few faulty connectors and transitions OR a few cause confusion OR faulty formatting or paragraphing	**4** Basic vocab or too much repetition; frequent errors of word choice, form, or colloc; frequent spelling errors; meaning may be confusing or unclear	**4** Limited or mostly simple structures; major problems in structures; frequent sentence structure, punctuation or capitalization errors; meaning sometimes confusing or unclear
Unsatisfactory	**2** TS unclear, faulty, or irrelevant; OR Opinion missing or inaccurate; OR Controlling idea inaccurate; -Zero if TS missing	**3** Reasons missing or too many; most details missing, unrelated, or unclear how examples/explanations relate to reasons or to TS -Zero if less than 3 lines/sentences.	**2** Concluding sentence /opinion unclear or irrelevant OR Summary inaccurate or missing -Zero if concluding sentence missing	**2** Very poor organization of ideas; connectors and transitions rare, repetitive, faulty, or missing; not enough to evaluate	Sentences are essentially **2** translated from Arabic; very limited control of word choice, form, or colloc; errors cause some strain for the reader; dominated by spelling errors; meaning is mostly unclear; not enough to evaluate	Very limited control **2** of sentence structure; dominated by punctuation or capitalization errors; errors may cause some strain for the reader; not enough to evaluate
0 if absent, paragraph plagiarized, or not submitted				**Score:** /50	**Average** /50 **Total**	/100 % (Pass ≥ 70%)

Figure 2

Writing Rubric – Opinion