APPLING
International Applied Linguistics Research Conference

https://doi.org/10.29140/9781763711686-5

Evaluating Large Language Models for Open-Ended Language Assessment: Rethinking Methodological Frameworks in Applied Linguistics

Darina Kallina
Independent Researcher, Russia

Abstract

Large language models (LLMs) are increasingly used to evaluate open-ended learner responses, but agreement with human raters alone does not establish validity. This paper proposes a multidimensional framework for evaluating LLM-based language assessment through a purposive synthesis of research in language assessment, automated writing evaluation, educational measurement, and generative AI scoring. The framework integrates four evidence sources: human–model agreement and stability, construct sensitivity, qualitative analysis of rationales and disagreements, and fairness across learner and task conditions. Russian as a foreign language (RFL) serves as an illustrative multilingual context. The synthesis suggests that strong aggregate agreement may coexist with prompt sensitivity, score instability, overreliance on surface linguistic features, and subgroup differences. The paper therefore argues for mixed quantitative and qualitative evaluation and transparent reporting of model version, prompt design, scoring repetitions, and disagreement cases. This framework supports more defensible investigation of when and for whom LLM-generated assessment judgements can be trusted.

Keywords: large language models; language assessment; automated writing evaluation; validity; Russian as a foreign language

E-mail: darina.kallina@gmail.com

In Y. Wang, J. Matthews, & B. Zou (Eds.), *Proceedings of the 2nd International Applied Linguistics Research Conference, 2026* (pp. 35–39). Castledown.

Background

Large language models (LLMs) are increasingly explored for evaluating open-ended learner responses. They can process extended texts, follow rubrics, and generate score explanations without the feature engineering required by earlier automated scoring systems. Studies report promising alignment with human judgements: Mizumoto and Eguchi (2023) found substantial agreement on non-native English essays, while Pack et al. (2024) reported strong reliability for GPT-4. Recent work likewise shows that LLM scoring can approximate expert judgement under some conditions (Crossley et al., 2026; Huang et al., 2026).

Agreement, however, is not equivalent to validity. Language assessment requires evidence that score interpretations and uses are defensible (Kane, 2013). An automated system may produce human-like scores for reasons unrelated to the intended construct, and earlier automated writing research has shown weaknesses in responding to higher-order features such as organisation and development (Chapelle et al., 2015).

The methodological problem is therefore broader than prediction accuracy. An LLM may correlate highly with human raters while remaining unstable near score boundaries, overvaluing grammatical accuracy, or reacting disproportionately to prompt wording. This paper develops a multidimensional framework for investigating such risks. RFL provides an illustrative context, while the framework remains language-independent.

Methodology

The study uses methodological synthesis rather than benchmark comparison, combining language assessment, automated writing evaluation, educational measurement, and generative AI research. The unit of analysis is the evidence needed to justify an LLM-generated assessment judgement.

Sources were selected purposively for direct relevance to the synthesis. Priority was given to peer-reviewed empirical and methodological studies on LLM scoring, automated writing evaluation, validity, reliability, fairness, and prompt sensitivity, supplemented by foundational validation research.

The framework combines four evidence sources. First, human–model agreement should be examined with design-appropriate indices such as correlations, adjacent agreement, weighted kappa, or intraclass correlation. Repeated scoring is also needed because average agreement can conceal model instability.

Second, construct sensitivity should test whether scores respond to features represented in the rating construct. Controlled perturbations are useful: weakening coherence while holding length and vocabulary stable should affect a construct-sensitive scorer, whereas superficial changes should not.

Third, scoring rationales and feedback should be analysed qualitatively for rubric alignment, specificity, and consistency. They can reveal which textual features the model treats as evidence and contradictions between scores and explanations, but should not be treated as direct access to the model's internal reasoning.

Fourth, fairness should be examined across learner profiles, proficiency levels, and task conditions. Subgroup comparisons should, where possible, control for relevant proficiency evidence so that genuine performance differences are not mistaken for bias (Chen et al., 2026). Prompt condition and repeated runs should also be considered.

Across all components, human–model disagreement should be analysed rather than discarded as noise. Boundary cases, large discrepancies, and judgements that change across prompts or runs can provide diagnostic evidence.

Representative Results

The synthesis shows that LLM performance cannot be represented adequately by a single agreement coefficient because different evidence sources reveal different failure modes and validity concerns.

Human–model agreement indicates whether an LLM broadly resembles expert scoring, but repeated runs may reveal hidden instability. Pack et al. (2024) observed scoring fluctuations, while Kim (2025) showed that prompting strategies affect reliability and alignment. Agreement is therefore partly a property of the scoring configuration.

Construct-sensitivity evidence distinguishes rubric-based evaluation from shortcut behaviour. A model may produce plausible holistic scores while relying on length, lexical sophistication, or surface correctness. Crossley et al. (2026) found that systems with similar agreement levels can relate differently to linguistic components and therefore represent the construct differently.

Qualitative disagreement analysis adds evidence about boundary effects, ambiguous criteria, and prompt-sensitive interpretations. If human raters reward communicative effectiveness while a model penalises comprehensible non-standard forms, construct representation may be at issue. Conversely, systematic disagreement can also expose weaknesses in human rubrics.

Fairness analysis may reveal subgroup-specific patterns hidden by overall accuracy. Automated approaches can vary across populations, proficiency

levels, and model configurations (Chen et al., 2026; Huang et al., 2026), making conditional subgroup analysis important in multilingual assessment.

Discussion

The framework shifts the question from whether an LLM can score responses to what evidence justifies trusting a judgement for a given purpose. This follows validity theory, where score use depends on a coherent argument rather than one technical indicator (Kane, 2013), while also addressing generative-model properties such as stochasticity and prompt sensitivity.

Researchers should report model version, interface or API, prompts, rubric wording, scoring repetitions, and examples included in prompts because these form part of the assessment instrument. Prompting choices can affect performance (Kim, 2025), and LLM raters differ in severity and reliability (Jiao et al., 2026).

For RFL and other multilingual settings, linguistic difference must be distinguished from construct-irrelevant penalty. A response may contain morphosyntactic errors while fulfilling a communicative task effectively. If an LLM systematically privileges surface accuracy, agreement on one dataset may conceal construct underrepresentation. For example, an RFL learner may communicate successfully while using non-target-like case endings or aspect choices; penalising these forms more heavily than the rubric warrants could underestimate communicative performance.

The framework does not prescribe universal thresholds because evidence requirements depend on assessment purpose and stakes. Classroom feedback, placement, certification, and research coding require different levels of justification. Human ratings also remain fallible, so disagreement should trigger investigation rather than automatic deference to either human or machine.

Conclusion

LLMs may become useful components of language assessment, but methodological credibility requires more than correlation with human ratings. Defensible evaluation should combine agreement and stability measures, construct-sensitivity tests, qualitative analysis of rationales and disagreements, and fairness checks. This multidimensional approach makes assumptions behind automated judgements visible and supports transparent investigation of when, how, and for whom LLM-generated scores can be considered valid.

References

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing, 32*(3), 385–405. https://doi.org/10.1177/0265532214565386

Chen, G., Wu, A. D., & Zhang, C. (2026). Assessing fairness in AI-assisted writing scoring: Developing fairness measures to detect predictive bias in automated essay scoring. *Assessing Writing, 69*, 101066. https://doi.org/10.1016/j.asw.2026.101066

Crossley, S., Holmes, L., & Morris, W. (2026). Assessing the reliability and validity of large language models in automatic essay scoring. *Assessing Writing, 69*, 101082. https://doi.org/10.1016/j.asw.2026.101082

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*, 101047. https://doi.org/10.1016/j.asw.2026.101047

Jiao, H., Song, D., & Lee, W.-C. (2026). Evaluating rater effects of large language models in automated essay scoring: GPT, Claude, Gemini, and DeepSeek. *Educational Measurement: Issues and Practice, 45*(2), e70018. https://doi.org/10.1111/emip.70018

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kim, Y. (2025). Automated essay scoring with GPT-4 for a local placement test: Investigating prompting strategies, intra-rater reliability, and alignment with human scores. *TESOL Quarterly, 59*(S1), S318–S329. https://doi.org/10.1002/tesq.3405

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), 100050. https://doi.org/10.1016/j.rmal.2023.100050

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*, 100234. https://doi.org/10.1016/j.caeai.2024.100234