QO‘QON DAVLAT PEDAGOGIKA INSTITUTI ILMIY XABARLARI
(2026-yil 10-son)
PEDAGOGIKA
PEDAGOGY
https://doi.org/10.70728/c.series.ped.v08.i10.061

FROM CEFR DESCRIPTORS TO AI-GENERATED FEEDBACK: RELIABILITY, VALIDITY, AND RESEARCH PRIORITIES FOR EFL WRITING ASSESSMENT IN UZBEKISTAN

Iskandar Vahobovich Rajabov
Asia International University,
The Department of Foreign Languages, Bukhara, Uzbekistan
E-mail: 35290034iskandar@oxu.uz

Abstract. Generative artificial intelligence (GenAI) is increasingly used to support English as a Foreign Language (EFL) writing through automated comments, language correction, and rubric-based scoring. Yet the usefulness of these applications depends on whether their outputs are consistent, aligned with intended assessment criteria, and pedagogically appropriate. This critical narrative review examines selected research on large language models in writing assessment and considers implications for CEFR-aligned assessment in Uzbek higher education. The discussion is organized around three issues: rating reliability, validity of score interpretations, and the quality and use of AI-generated feedback. Existing evidence indicates that large language models can produce consistent rubric-based ratings under particular conditions, but consistency alone does not establish that a score validly represents writing proficiency or a CEFR level. Feedback may support revision, particularly when it addresses language form and textual organization, but learners and teachers still need to evaluate the relevance and accuracy of suggestions. The review identifies a context-specific research priority: comparing AI-generated judgments with trained human ratings using transparent CEFR-informed criteria and investigating how learners use the resulting feedback. Because this article is a critical narrative review rather than a systematic review or an original empirical study, its conclusions are provisional and should be followed by a documented literature search and local validation study.

Keywords: generative artificial intelligence; CEFR; EFL writing assessment; automated feedback; reliability; validity; Uzbekistan

1. INTRODUCTION. Writing assessment plays an important role in English language education because it provides evidence of learners’ ability to communicate meaning, organize ideas, select appropriate vocabulary, and exercise grammatical control. In higher education,

assessment results can inform instructional planning, provide learners with information about progress, and contribute to decisions about achievement. However, evaluating written work carefully requires time and consistent application of assessment criteria. Generative artificial intelligence (GenAI), especially large language models (LLMs), has consequently attracted attention as a possible source of rapid feedback and rubric-based scoring.

LLMs can identify possible language errors, propose revisions, explain grammatical patterns, and generate comments on organization or task fulfilment. These functions may make formative feedback more accessible between lessons. Nevertheless, fluent and confident output should not be equated with accurate assessment. A model may overlook a problem, recommend a change that alters the writer’s intended meaning, or assign a plausible score without adequately applying the intended construct.

The Common European Framework of Reference for Languages (CEFR) provides a shared framework of illustrative descriptors for describing language proficiency. The Companion Volume expands and updates these descriptors, supporting their use in teaching, learning, and assessment (Council of Europe, 2020). Applying CEFR descriptors is not simply a matter of asking an AI system to assign a level. A defensible judgment requires a clear relationship between the writing task, observable evidence in the text, the relevant descriptors, and the interpretation made from the resulting score.

Research has started to examine the reliability and validity of LLM-based essay grading. Yavuz et al. (2025), for example, compared human EFL instructors’ ratings with ratings generated by ChatGPT and Bard using an analytic rubric. Chuang and Yan (2025) synthesized 77 studies on GenAI in language testing and assessment and emphasized that reliability, validity, fairness, and practicality should be considered together. Research on AI-assisted feedback also suggests potential benefits for revision, while indicating that effects depend on the learning design and context (Mekheimer, 2025).

The purpose of this article is to critically discuss what these strands of research imply for CEFR-aligned EFL writing assessment in Uzbekistan. It addresses three questions: (1) What does selected research indicate about the reliability of LLM-based writing assessment? (2) What validity issues arise when AI-generated judgments are interpreted in relation to CEFR descriptors? (3) What research priorities should guide the responsible use of AI-generated feedback in Uzbek higher education?

2. Literature Review

2.1. Generative AI and writing assessment

Automated writing evaluation includes technologies that assess written texts or provide feedback to support revision. Earlier systems often relied on predefined linguistic features or statistical patterns. LLMs add the ability to produce contextualized natural-language explanations and respond to follow-up questions. This flexibility can help learners ask why a

phrase is unclear, compare alternative formulations, or request feedback on a specific rubric criterion.

Yet the apparent detail of AI-generated comments can obscure their limitations. Feedback may be linguistically plausible but not appropriate for the task, genre, or intended meaning. A system may also focus on surface-level correctness because these features are easier to identify than the quality of an argument, originality, audience awareness, or successful fulfilment of a communicative purpose. Therefore, the quality of feedback should be evaluated against explicit criteria rather than judged by fluency alone.

Fairness is closely related to validity. A model may respond differently to texts containing unfamiliar names, culturally specific examples, non-standard but intelligible expressions, or language patterns associated with particular learner groups. Such differences should not automatically be interpreted as differences in proficiency. For this reason, studies should examine examples of disagreement and consider whether the model penalizes features that are not part of the intended construct.

Prompt design is also part of the assessment procedure. Instructions that specify a rubric, require evidence for each criterion, and discourage unsupported inferences may influence the form and consistency of model outputs. However, a detailed prompt is not a substitute for validation. Researchers should document the exact prompt and model version so that the procedure can be scrutinized and, where possible, replicated. Changes to a commercial model over time may affect reproducibility even when the prompt remains unchanged.

A further distinction is needed between inter-rater agreement and repeatability. Repeatability concerns whether a system produces similar outputs when the same material is assessed more than once under comparable conditions. Inter-rater agreement concerns the degree to which different assessors, such as an AI model and trained teachers, assign sufficiently similar ratings. These questions require different designs and statistics. A study that reports only a correlation between AI and teacher scores may overstate agreement because correlation does not establish that the scores are close enough for interchangeable use.

2.2. Reliability and validity

Reliability concerns the consistency of scores under specified conditions. For LLM-based assessment, researchers can examine whether repeated ratings of the same text remain stable and whether AI ratings agree with ratings from trained human assessors. Yavuz et al. (2025) reported high intraclass correlation coefficients for the models and conditions examined: 0.947 for default ChatGPT, 0.972 for a more tightly prompted ChatGPT configuration, and 0.919 for Bard. The study used only three essays, however, so its findings should not be generalized uncritically to other tasks, models, proficiency levels, or institutional settings.

Validity concerns the interpretation and use of assessment results. A system could produce similar scores repeatedly while consistently overlooking a relevant aspect of writing.

Likewise, agreement with a human assessor does not automatically establish that either judgment adequately represents the intended construct. A CEFR-related score should be supported by evidence that the criteria, task, and rating process represent the proficiency interpretation being claimed. This distinction is especially important if an AI score may affect grades, placement, or certification.

2.3. CEFR descriptors and feedback literacy

The CEFR is a reference framework, not a single universal scoring rubric. Descriptors need to be selected and interpreted in relation to the task and the purpose of assessment. A model prompted to label an essay B1 or B2 may produce a plausible label, but that output should not be treated as a validated proficiency judgment without evidence of alignment and appropriate comparison with expert ratings. Teachers should be able to identify which features of the writing support a judgment and where uncertainty remains.

Feedback is most educationally useful when learners can understand it, evaluate its relevance, and apply it to revision. AI can increase the availability of comments, but the number of comments does not guarantee learning. Learners may accept incorrect suggestions or make local corrections without understanding the underlying principle. Teacher guidance can help learners prioritize revisions, question questionable feedback, and connect comments to assessment criteria. Consequently, AI feedback literacy—the ability to interpret and critically evaluate AI-generated feedback—deserves attention alongside the technical performance of the model.

Language assessment also involves practical constraints. Teachers need procedures that can be implemented within available time, training, and technology. An approach that is statistically promising but difficult to explain to students or impossible to audit may have limited educational value. A locally useful framework should therefore balance technical performance with transparency, data protection, teacher workload, and learners’ opportunity to question feedback.

Another methodological issue concerns the meaning of a CEFR level in a particular institution. A course may use CEFR descriptors as a broad learning reference while its local examination uses a separate rubric or score scale. These systems may overlap without being equivalent. Before AI ratings are described as CEFR-aligned, researchers should specify which descriptors are relevant, how they have been operationalized, and whether the task elicits the performance those descriptors describe.

2.4. The Uzbek higher-education context

International findings cannot be assumed to transfer directly to Uzbek higher education. Local institutions may use different writing tasks, proficiency expectations, assessment rubrics, and instructional practices. Learners’ familiarity with AI tools and teachers’ experience in applying CEFR descriptors may also vary. The key issue is therefore not whether AI can generate a CEFR label in general, but whether AI-generated feedback and

scores provide useful and defensible evidence in the local assessment context. This article treats that issue as a research priority rather than claiming that the gap has already been conclusively established by an exhaustive review.

3. Methodology

This article uses a critical narrative review approach. It synthesizes selected publications that directly address LLM-based writing assessment, language-assessment principles, AI-generated feedback, and the CEFR framework. The principal sources include the Council of Europe’s CEFR Companion Volume (2020), the empirical study by Yavuz et al. (2025), the systematic review by Chuang and Yan (2025), and research on GenAI-assisted EFL writing feedback by Mekheimer (2025).

The sources were considered in relation to three analytical themes: (a) consistency of AI-generated ratings, (b) alignment between scores and the intended assessment construct, and (c) pedagogical usefulness of feedback. Findings reported by the original studies are distinguished from implications proposed for Uzbek higher education. The article does not claim to be a systematic review: it does not report an exhaustive database search, preregistered protocol, formal screening flow, or comprehensive risk-of-bias assessment. A systematic search and documented screening process would be necessary for a study making claims about the complete body of available evidence.

Across the selected sources, the central pattern is that the promise of AI-based writing assessment is accompanied by a need to define the conditions under which its outputs can be trusted. The evidence does not support a universal claim that generative AI is either a reliable replacement for teachers or an unsuitable tool for all assessment purposes. Rather, the defensibility of its use depends on the task, rubric, model, population, and consequences attached to the score.

4. Results of the Critical Synthesis

4.1. AI ratings can be consistent under specific conditions

The selected literature indicates that LLMs can produce consistent ratings when prompts and scoring criteria are specified. The reliability coefficients reported by Yavuz et al. (2025) illustrate that repeated AI ratings can be evaluated with established methods. However, the study’s small set of three essays limits the range of writing performances represented. Reliability should therefore be examined across a larger and more varied sample before an AI system is relied on for broader assessment decisions.

4.2. Consistency does not establish CEFR validity

A second theme is the need to distinguish stable scoring from defensible score interpretation. CEFR alignment requires more than a prompt containing CEFR terminology. Researchers should investigate whether the system applies relevant descriptors to observable features of the text, whether its criterion-level judgments agree with trained raters, and whether disagreements are systematic. Chuang and Yan (2025) similarly emphasize that

validity, fairness, and practicality must be considered alongside reliability in GenAI-based language assessment.

4.3. Feedback needs pedagogical and human review

The literature also suggests that AI feedback should be evaluated as a learning resource rather than assumed to be beneficial because it is immediate. Feedback may support revision, but its usefulness depends on accuracy, relevance, clarity, and learners' ability to apply it. A responsible classroom model would treat AI comments as provisional: learners consider the suggestions, teachers verify important judgments, and assessment decisions remain transparent and accountable.

Table 1

Analytical themes and implications for CEFR-aligned writing assessment

Theme	What the selected literature indicates	Implication for Uzbekistan
Reliability	LLMs can show stable rubric-based ratings under particular conditions.	Test repeated ratings and human–AI agreement across a larger, varied sample.
Validity	Consistent scores do not by themselves prove valid interpretation of proficiency.	Map judgments to selected CEFR descriptors and examine criterion-level disagreements.
Feedback use	Immediate feedback may support revision, but relevance and uptake matter.	Study whether learners evaluate and use AI feedback accurately, with teacher oversight.

The proposed research agenda can be developed progressively. An initial pilot can investigate the agreement between AI and trained human raters on a carefully selected set of anonymized writing samples. A second study can focus on the accuracy and usefulness of criterion-specific feedback. A third study can investigate learners' engagement with that feedback and the quality of subsequent revisions. Only after these stages provide sufficient evidence would it be appropriate to propose a broader institutional model for human–AI collaboration. This sequence reduces the risk of making large claims from a small or unvalidated pilot.

Ethical and governance questions should also be incorporated into the research design. Student writing should be anonymized before being entered into external AI systems, and institutional requirements for consent and data protection should be followed. Learners should be told whether AI is used to generate formative comments or contribute to a grade. If an AI output influences a consequential decision, students should have access to human review and a clear route for challenging an assessment judgment. These safeguards are not optional additions; they affect the transparency and acceptability of the assessment process.

The choice of statistical analysis should follow the research question. For score agreement, researchers may consider an intraclass correlation coefficient or a weighted kappa statistic where appropriate to the score scale and design. They should report confidence

intervals where feasible and describe the distribution of disagreements. For feedback quality, a coding scheme could classify comments as accurate, partially accurate, irrelevant, misleading, or insufficiently explained. Such categories should be defined before coding, and a second researcher or trained coder can be used to check the consistency of the coding process.

A practical framework for future research can be organized into four stages. First, researchers should define the assessment purpose and identify the specific CEFR descriptors that are relevant to the writing task. Second, they should prepare a transparent rubric and train human raters to apply it, checking agreement before comparing their ratings with AI outputs. Third, the AI procedure should be documented, including the model, prompt, input format, and whether repeated ratings are permitted. Fourth, discrepancies should be analyzed at both score and comment levels rather than summarized only as an overall agreement coefficient.

5. Discussion

The synthesis suggests that generative AI should be treated as a potential component of assessment practice, not as an automatically dependable assessor. Three implications follow. First, reliability studies should specify the model and version, prompt, rubric, task, scoring conditions, and number of repeated ratings. When comparing AI with human assessors, researchers should select appropriate agreement statistics and report uncertainty. Correlation alone is insufficient because two sets of scores can correlate strongly while differing systematically in level.

Second, validity claims should be proportionate to the evidence. A model that provides useful comments on grammar may not necessarily judge argument quality, task fulfilment, or overall proficiency accurately. Studies should report results at both overall-score and criterion levels, investigate systematic disagreements, and avoid equating a generated CEFR label with an independently validated proficiency classification.

Third, the role of feedback in learning should be studied separately from the accuracy of scoring. A useful research design could examine whether learners understand AI comments, which suggestions they accept or reject, and whether revisions improve according to a clearly defined rubric. Such a study should avoid assuming that all revisions are beneficial; learners may introduce new errors or alter meaning while following a suggestion.

For Uzbekistan, a feasible next step is a small-scale validation study using anonymized student writing samples, a transparent rubric linked to selected CEFR descriptors, and at least two trained human raters. AI-generated scores and comments could then be compared with human judgments at the criterion level. The study should document consent, data protection, the model used, and the procedure for resolving disagreements. If the research later investigates learning effects, a separate design could compare revisions made with AI feedback, teacher feedback, or a structured combination of both.

This agenda offers a more defensible route to a locally relevant contribution than making broad claims that AI improves writing or that it can replace teachers. It also creates a

connected research programme: first evaluate rating agreement, then examine feedback quality and learner uptake, and finally develop and validate a human–AI assessment protocol. The present review does not provide local empirical evidence, so these proposals should be understood as priorities for future research rather than findings about Uzbek students.

Several limitations should be acknowledged. The article is a critical narrative review and does not present a reproducible systematic search or a new dataset. The sources discussed are selective and cannot represent every available study on AI writing assessment. In addition, AI systems change frequently, meaning that findings for one model version may not transfer to another. These limitations restrict the strength of the conclusions and reinforce the need for a documented review and local empirical validation.

6. Conclusion

Generative AI can provide rapid comments and rubric-based ratings for EFL writing, but its responsible use requires more than fluent feedback or stable scores. Reliability, validity, fairness, and pedagogical usefulness are related but distinct concerns. In particular, consistency in scoring does not establish that an AI-generated judgment validly represents writing proficiency or a CEFR level.

For Uzbek higher education, an important next step is to compare AI-generated ratings and feedback with judgments from trained human assessors using transparent CEFR-informed criteria. Future research should investigate criterion-level agreement, the accuracy and relevance of feedback, and learners’ ability to use AI comments critically. This article is a preliminary critical narrative review and does not claim exhaustive coverage of the literature. Its contribution is to define a focused research agenda that can be tested empirically and developed into a locally validated approach to human–AI collaboration in language assessment.

REFERENCES

1. Chuang, P.-L., & Yan, X. (2025). Language assessment in the era of generative artificial intelligence: Opportunities, challenges, and future directions. System, 134, Article 103846. https://doi.org/10.1016/j.system.2025.103846

2. Council of Europe. (2020). Common European framework of reference for languages: Learning, teaching, assessment—Companion volume. Council of Europe Publishing. https://www.coe.int/en/web/common-european-framework-reference-languages/cefr-companion-volume-and-its-language-versions

3. Mekheimer, M. (2025). Generative AI-assisted feedback and EFL writing: A study on proficiency, revision frequency and writing quality. Discover Education, 4, Article 170. https://doi.org/10.1007/s44217-025-00602-7

4. 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. https://doi.org/10.1111/bjet.13494