Journal of Artificial Intelligence and Technology, 2026, 6, 349-355
https://doi.org/10.37965/jait.2026.0949

ISTP
RESEARCH ARTICLE

Evaluating Large Language Models for Educational Measurement Insights from Automated and Human Scoring of Language Exams

Bora Başaran
Department of Foreign Languages Education, Anadolu University, Eskişehir, Türkiye

(Received 26 September 2025; Revised 07 January 2026; Accepted 27 January 2026; Published online 22 February 2026)

Abstract: This study investigates the use of large language models (LLMs)—ChatGPT-5, Claude Opus 4.1, Gemini Advanced 2.5 Pro, DeepSeek Pro, Qwen-3 Max, and Mistral Le Chat Pro—and a locally fine-tuned LLaMA 3.3 70B Instruct model for automating assessment tasks in language education. Specifically, the study looks to examine LLM capabilities in automating assessments with authentic midterm exam sheets from a "German as a Foreign Language" (GFL) course in three different scenarios: (1) general-purpose LLM with pre-corrected samples giving a score, (2) localized grading using a fine-tuned LLaMA model and reference answer keys, and (3) manual grading with and without a visual overlay technique. Human grading, supported by a structured scoring process, remained nearly perfect in terms of accuracy and reliability, whereas the local model failed because OCR and visual input techniques did not produce usable outputs. These findings reinforce the necessity for domain-specific adaptation, the design of stronger OCR and multimodal workflows, and explainable scoring mechanisms before local AI solutions can reliably contribute to the assessment of language learning tasks in applied educational settings.

Keywords: Artificial intelligence in education; automated assessment; educational technology; human–AI comparison; language teaching; large language models (LLMs)

I. INTRODUCTION

The educational scene is transforming substantially because of AI technologies that are redefining classroom methods [1–5]. In language education, the potential and limitations of AI are pronounced due to the intricacies of language acquisition and assessment [6–9]. Because of the rapid adoption of large language models (LLMs) (ChatGPT, Claude, etc.), researchers and scholars are investigating various ways to automate assessment to lighten the load of assessment and provide faster feedback [10–12]. However, empirical evaluation in this area remains preliminary [13,14].

German presents particular challenges for assessment due to its case system, grammatical gender, and extensive use of compounding. We thus examine general-purpose LLMs and a fine-tuned local model on authentic exam papers in a German teacher education program and ask:

RQ1. How accurately can general-purpose LLMs evaluate in-program exams when given pre-corrected samples as references?
RQ2. How well does a fine-tuned local model score student responses against reference answer lists?
RQ3. How do AI-based techniques compare with traditional human grading, including a newly developed overlay method?
RQ4. Which integration strategies balance accuracy, efficiency, and usability in test assessment?

The remainder of this paper is structured as follows. Section II reviews prior research on technology-assisted language assessment, LLMs in education, and fine-tuning approaches for domain-specific evaluation tasks, with particular attention to German language assessment. Section III describes the research design, dataset, assessment conditions, and analytical procedures. Section IV presents the empirical results addressing accuracy, reliability, efficiency, and usability across human and AI-based grading methods. Section V discusses the findings in relation to methodological constraints and pedagogical implications. Finally, Section VI concludes the paper by summarizing key contributions, limitations, and directions for future research.

II. RELATED LITERATURE

A. EVOLUTION OF TECHNOLOGY IN LANGUAGE ASSESSMENT

Language education technology has transitioned over the last three decades, from early behaviorist Computer-Assisted Language Learning (CALL) to engaging, data-informed, and immersive platforms [15–21]. AI made ASR-based feedback and adaptive platforms possible [22], but it has also come with concerns over data privacy, equity, and teacher readiness and willingness to change [23–25]. Strength of evidence indicates that hybrid approaches, where human pedagogy is combined with technology, are preferable [16,21,25,26].

B. LARGE LANGUAGE MODELS AND THEIR EDUCATIONAL APPLICATIONS

AI is reshaping instruction and assessment through personalization and feedback speed [27–32]. Assessment using LLMs has shown promise but lacked coherence, cultural context, and complex domain-specific tasks [33–39]. Performance is language-specific due to dataset differences; non-English and minority languages

Corresponding author: Bora Başaran; (e-mail: bbasaran@anadolu.edu.tr).

tend to not perform as well [35,40,41]. German had intentions—results are mixed and task-dependent [42–44].

C. FINE-TUNING FOR SPECIALIZED EDUCATIONAL TASKS

Fine-tuning has the potential to enhance task alignment and efficiency with small domain datasets [45–47]. Curriculum/domain adaptation and adapter-based methods provide opportunities for resource constraints [48,49] and have reported improvements in educational scoring tasks [50,51]. Overall, practical implementation of fine-tuning is dependent on a reliable data pipeline (OCR/multimodal), interface input, and input data quality—voids that can offset fine-tuning benefits theoretically [52,53].

D. ASSESSMENT IN GERMAN LANGUAGE TEACHING

German has a particularly complicated structural intricacy, with its case system, gendered articles, and inflectional morphology of words, which makes it especially challenging with respect to instruction and assessment [54,55]. These complexities often add cognitive load and complicate the assessment of learners' writing [56]. Although standardized DaF exams test multiple competencies, that is not true for language teacher education at the university level, which often prioritizes open-ended, information-rich exams [57]. Recent developments in automated assessment (e.g., G-SciEdBERT) have aimed to help teachers to assess the German language [43,58]. However, results have been mixed, especially for complex and natural writing. This work is not about standardized DaF tests or grammar assessment schemes—this is about using a general or fine-tuned LLM for assessing authentic midterm assessments in German teacher education. This study also extends the literature that is developing as it addresses technical and pedagogical shifts in the deployment of AI in classroom-based language assessment [59–62].

Fine-tuning pretrained models on narrower and specific datasets can enhance their capability and performance for specific tasks and contexts (e.g., [45]). This offers flexibility and opportunity for tuning powerful generic models that would cost far more to operationalize based on scratch. For example, fine-tuning language models for education improves both the efficiency and contextual relevance of AI-based solutions. Fine-tuning LLMs for educational context enhances both efficiency and contextual relevance of AI-based solutions, particularly for automated assessment and feedback [45]. This is illustrated through the assessment of STEM education with GPT-3.5, where results indicate 9.1% better accuracy compared to baseline models [51]. The evidence suggests that fine-tuning is most beneficial when language and contextual details are critical of the domain. Suggestions for curriculum domain adaptation and adapter-based fine-tuning address the issues of client resources and overfitting [63]. The suggestions balance work efficiency through data difficulty ranking with unsupervised adaptation or modular trainable layers that maintain general knowledge and tweak networks' behavior [48,49]. The fine-tuning method is of utility in many fields, enhancing language models for machine translation, information extraction, cross-linguistic performance, and a wide range of tasks [64,65]. Assessment modeling in interactive learning systems involves, for example, modeling assessment that predicts a pedagogical assessment based on learner–system interaction data as a pretraining task for improving adaptiveness and reduce dependence on labeled data [66]. General and domain-specific model efficacies were balanced based on the data curation and modeling standards, as well as systematic evaluation and assessment of model architecture and learning development/evolution [67]. Thus, the fine-tuning process provides a flexible and extensible basis for harmonizing AI tools about the complicated context of evaluation in education. The field testing of such applications is especially challenging in limited-resource locations. The current study demonstrates that local fine-tuning models do not yield reliable, usable results without technical support and stability of the input pipelines (e.g., OCR) and accurate, high-quality annotated training data. The theoretical potential of fine-tuning in education needs ongoing infrastructure support and sustained supervising and controlled contextual evaluation to become practical (e.g., [52,53,68]).

III. METHODOLOGY AND PROCEDURES

A. RESEARCH DESIGN

Using a comparative experimental design with the Faculty of Education at Anadolu University, we have compared (1) automated scoring with general-purpose LLMs, (2) an overall locally fine-tuned (LLaMA-based) model, and (3) human grading with regular keys and with an overlay method.

B. Dataset

Ten actual anonymized midterm response papers from "Approaches to Foreign Language Teaching" (more than 50% multiple-choice items) are used. All PII and ethical permissions were submitted and obtained per institutional ethics. AI-based approaches are presented structured inputs through standardized answer keys provided by the university. General-purpose LLM performance was moderate; the local model was limited by the interface and poor visual/OCR, resulting in limits to comparability.

The dataset size was intentionally limited to 10 exam papers because this study was designed as an exploratory pilot investigation aimed at identifying comparative performance patterns rather than producing statistically generalizable claims. The selected papers were drawn from fully completed and legible midterm exams and were considered representative of typical student performance within the course. This controlled sample enabled a detailed, methodologically transparent comparison of human and AI-based scoring behaviors under identical assessment conditions.

Although more than half of the exam consisted of multiple-choice items, these items were not limited to surface-level factual recall. Instead, they required learners to process morpho-syntactic structures, grammatical agreement, and contextual interpretation, which are central challenges in German as a foreign language. Consequently, the assessment construct targeted in this study encompasses linguistic decision-making processes embedded in authentic exam conditions rather than exclusively open-ended writing performance.

C. ASSESSMENT APPROACHES

General-purpose LLMs are ChatGPT-5, Claude Opus 4.1, Gemini Advanced 2.5 Pro, DeepSeek Pro, Qwen-3 Max, and Mistral Le Chat Pro. The inputs include (a) a scan of the answer sheet, (b) the answer key/list of textual reference answers, and (c) student sheets, both graded and ungraded (text/image). The grading methods include (1) task + criteria, (2) criteria + graded examples with

full explanations, and (3) a collection of ungraded student responses. No explicit programming of rubrics is used; instead, we stimulate pattern recognition and contextual reasoning.

The fine-tuned local model, LLaMA 3.3 70B Instruct, was trained on a set of 142 graded midterms (LM Studio v0.3.14; 10 epochs; lr = 5e-5; batch size = 16; AdamW) [69–71]. At inference, the model received keys and returned intended scores and rationale. At the time of writing this (i.e., two months after reviewing), the model did not output usable data due to OCR/visual parsing and interface limitations.

This condition was intentionally retained in the study to evaluate the practical feasibility of deploying a locally fine-tuned LLM in an authentic, resource-constrained educational setting. Accordingly, the fine-tuned LLaMA condition functions as a feasibility and stress-test case rather than as a fully comparable scoring condition, allowing the study to empirically document infrastructural and pipeline-related barriers that may undermine the benefits of fine-tuning in real-world assessment scenarios.

Human grading. Two experienced DaF instructors graded independently using the answer keys and rubrics (tracked time). The overlay method involved corresponding intervals with publishers' transparent layers with correct selections for visual centering in case of concision and visual/recognition accuracy (ultimately prioritizing scoring accuracy and usability over precise timing measurements).

IV. DATA COLLECTION AND ANALYSIS

The assessment performance evaluation examined three fundamental elements which included accuracy alongside efficiency and reliability. Accuracy: Each evaluation method received accuracy assessments through comparison with a gold standard score that two experienced instructors generated together. The "gold standard" represents a widely recognized benchmark that serves as a reference point for comparative validation in educational assessment research [72,73]. Accuracy metrics included:

- The comparison between each evaluation method and the gold standard score used percentage agreement as its metric.
- Cohen's kappa coefficient to assess interrater reliability [74,75].

For the purposes of this study, Cohen's κ was calculated at the total-score level after scores were mapped to discrete agreement categories, enabling categorical comparison between each AI-based scoring output and the human gold standard. This approach was adopted to assess systematic agreement patterns rather than to model continuous score deviation, which was separately captured through root mean square error (RMSE).

- RMSE calculated the mean distance between predicted scores and gold standard values.

RMSE functions as a common evaluation metric for models because it works best with Gaussian error distributions. The squared error component in RMSE enhances its sensitivity to large deviations while increasing penalties for such deviations [76–78]. Efficiency and reliability: The evaluation time required to examine each paper served as the main efficiency measurement. The total evaluation time for AI-based methods consisted of processing time for model output generation and human intervention time for input formatting and result interpretation. Human evaluators directly assessed papers without needing additional methods. The overlay method sought to expedite evaluation through visual alignment but did not obtain any formal timing data. The institution paid human raters for their complete time spent on each paper according to standard research participation compensation rules.

V. RESULTS

This section outlines the findings from the study that examined research questions RQ1–RQ4 that related to accuracy, efficiency, and usability of evaluation approaches that incorporated AI and human evaluations of German language examinations.

RQ1: How accurately can general-purpose LLMs evaluate in-program examinations in a German teacher education context by comparing them to pre-corrected samples? Accuracy was evaluated against a gold standard constructed by two human evaluators with expertise in research on German language pedagogy. All RMSE values reported in this study were calculated based on total exam scores, with the maximum achievable score set at 100 points. RMSE therefore represents the average absolute deviation of each scoring method from the human gold standard on the full-exam scale, allowing direct interpretability of error magnitude across human and AI-based evaluation approaches. Against the human gold standard, we obtained (see Figs 1 and 2):

- ChatGPT-5: RMSE = 5.000, $\kappa = 0.344$
- Claude Opus 4.1: RMSE = 6.480, $\kappa = 0.249$
- Gemini Adv. 2.5 Pro: RMSE = 7.424, $\kappa = 0.028$
- DeepSeek Pro: RMSE = 8.485, $\kappa = 0.033$
- Qwen-3 Max: RMSE = 4.690, $\kappa = -0.018$
- Mistral Le Chat Pro: RMSE = 5.385, $\kappa = -0.164$

Percentage agreement complements these metrics (see Table I):

Human grading achieved 98% accuracy. Among the LLMs, Qwen 3 Max (78%) performed best, followed by ChatGPT-5 (75%) and Mistral (71%). However, higher agreement does not always imply higher κ; negative κ values indicate systematic deviation from human decisions.

The performance of ChatGPT-5 metrics for output is the most balanced overall, while Qwen-3 Max had the highest percent agreement at 78%; ChatGPT-5 is next at 75%, followed by Mistral at 71%. The models with high percent agreement show inconsistency in the same response alignments among question types and have low interrater agreement reliability. Some of the models with negative kappa values indicate that they systematically disagreed with the human gold standard. Ideally, the LLMs sometimes come up with the right answer, yet collectively their measure of reliability was limited overall.

Table I. Accuracy metrics of AI and human grading compared to reference grading

Evaluation method	Percentage agreement (%)	RMSE	Cohen's kappa
Human grading	98	1.414	0.926
Qwen-3 Max	78	4.690	−0.018
ChatGPT 5 grading	75	5.000	0.344
Mistral Le Chat Pro	71	5.385	−0.164
Claude Opus 4.1	58	6.480	0.249
Gemini Adv. 2.5 Pro	45	7.424	0.028
DeepSeek Pro grading	28	8.485	0.033

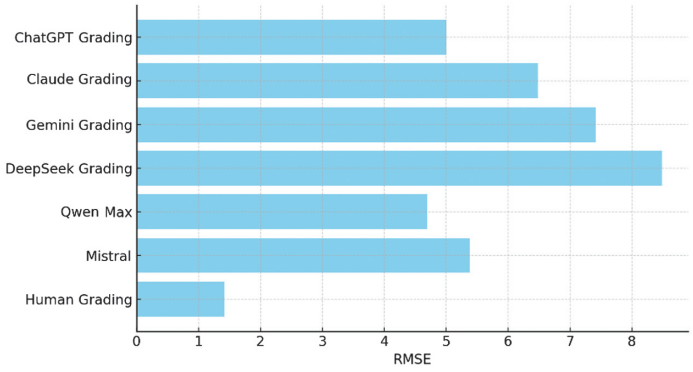

Fig. 1. RMSE of assessment methods compared to reference grading.

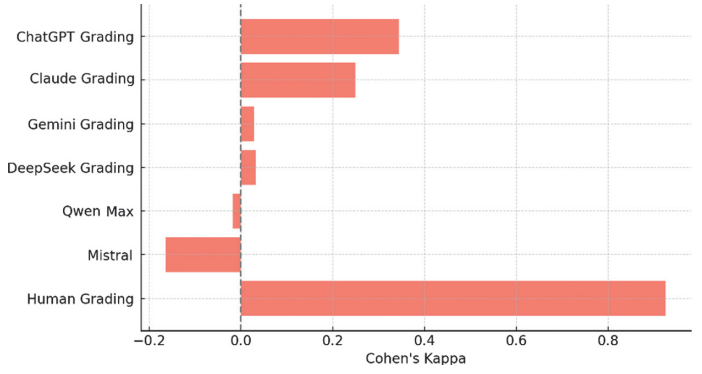

Fig. 2. Cohen's kappa of assessment methods compared to reference grading.

RQ2—Fine-Tuned Local Model
The local LLaMA 3.3 70B model fails to complete scoring; RMSE/κ cannot be computed. Primary causes are absent OCR integration, unstable visual parsing, and LM Studio interface limitations. Fine-tuning alone is insufficient without a robust multimodal pipeline. As a result, Research Question 2 could not be answered empirically in terms of quantitative accuracy or reliability metrics. The absence of valid model outputs precluded statistical comparison with human or general-purpose LLM scoring results, and the findings related to RQ2 are therefore restricted to feasibility-level observations rather than performance-based evaluation.
RQ3—Human vs. AI Methods
Human methods are most reliable: RMSE = 1.414, $\kappa = 0.926$. The overlay technique matched conventional accuracy and—based on qualitative observation—reduced cognitive load and sped up operations (though not formally timed).
RQ4—Integration, Efficiency, Usability

- Approximate end-to-end times per paper:
- Human (conventional): ∼3.5 min
- LLMs: 15–30 s model latency, but ∼2–3 min including prompt prep and verification
- Overlay: fastest human method (not formally timed)
- Fine-tuned LLaMA: >30 min with no valid output

Usability: Overlay = intuitive/low error; LLMs = verification needed; local model = unworkable under current constraints.

VI. DISCUSSION

A. COMPARATIVE EFFECTIVENESS OF ASSESSMENT APPROACHES

Results confirm prior work: educational AI requires domain-specific adaptation and stable input pipelines [79–83]. Despite moderate performance in places (e.g., ChatGPT-5 $\kappa = 0.344$), LLMs were unreliable on linguistically nuanced items. The local model illustrates deployment barriers: absent OCR/multimodal support can nullify fine-tuning gains. Human grading remains the most dependable approach; the overlay method is a pragmatic enhancement that preserves accuracy while improving speed and assessor comfort. In the short term, human-in-the-loop workflows are the most sensible path. Importantly, the fine-tuned local LLaMA condition should not be interpreted as evidence against the theoretical value of fine-tuning itself. Rather, the failure to obtain usable outputs highlights the dependency of fine-tuned models on stable OCR, multimodal input pipelines, and interface reliability. In this respect, the RQ2 findings contribute methodological insight by demonstrating that infrastructural limitations can invalidate otherwise well-aligned fine-tuning strategies in applied educational assessment contexts.

B. PRACTICAL IMPLICATIONS FOR LANGUAGE TEACHING

In practice, general-use LLMs should primarily be used for formative purposes (e.g., quick screening and draft feedback). High-stakes and summative grading should not rely on LLMs without human supervision, they should not use an LLM [84]. When institutions explore fine-tuned models [85,86], the best starting points are to invest in tested OCR and multimodal input pipelines, sturdy and reliable interfaces, and the development of high-quality, annotated datasets. To offload work but maintain some fidelity in multiple-choice-heavy environments, the overlay templates should be normalized at the institutional level and shared across courses. Finally, requisite programs should determine explicit human verification points and stipulated guidelines that ensure the validity, fairness, and teacher's pedagogical integrity are ensured throughout the assessment process.

C. FUTURE DIRECTIONS FOR AI IN LANGUAGE ASSESSMENT

Future research should focus on explainable AI (XAI) scoring rationales that educators and stakeholders can build trust in and that can help identify mistakes. A combination of AI and human raters should be utilized: AI can score objective items, while human raters evaluate subjective or multifaceted responses. The optimal form of hybridization should be determined based on empirical evidence [38]. It is also necessary to produce customizable, teacher-facing tools for non-programmers to help localize models to their curriculum and assessment style. Finally, we need to ramp up multilingual training and evaluation protocols to mitigate data volatility and cultural misalignment in order to establish reliability of these non-English contexts.

VII. CONCLUSION

This study demonstrated that, while AI-based assessment methods provided efficiency gains, they did not achieve the level of

grammatical sensitivity, syntactic awareness, and linguistic nuance required for valid language assessment. General-purpose LLMs exhibited inconsistent performance, and the fine-tuned local model did not yield usable results due to infrastructural limitations in OCR and multimodal processing. In contrast, human grading—particularly when supported by the overlay technique—achieved the most reliable balance between accuracy and efficiency. These findings underscore that, under current technological conditions, human-centered and hybrid assessment systems represent the most justifiable and methodologically sound approach for evaluating language learning outcomes.

VIII. RECOMMENDATIONS

Utilize a hybrid assessment: use LLMs to pre-assess or mark structured items with human involvement; use trained instructors for subjective/open-ended items. Invest in an input path with OCR for scanned/handwritten documents, and also take advantage of the XAI component to justify assessments and reveal mistakes. Train teachers with technology for ethical and efficient AI use, and allow for low-tech human augmentation (e.g., overlay) as a scalable, cost-effective enhancement.

ETHICS CONSIDERATIONS

This study is approved by the institutional review board at Anadolu University. Data were anonymized; participation was covered via exam-response use for research. There was no commercial influence. Potential AI bias and transparency limitations were considered in interpretation.

LIMITATIONS

A small dataset (10 papers; modest fine-tuning corpus) limits generalizability. The local model lacked OCR/multimodal stability. LLMs were not trained on education-specific corpora, affecting contextual precision. Human grading, while strong, still carries subjective variability. The scope focused on multiple-choice items; future work should evaluate open-ended responses. In addition, the small sample size limits statistical power and precludes inferential testing or generalization beyond the studied context. Accordingly, all reported accuracy and reliability metrics should be interpreted as descriptive indicators rather than as confirmatory evidence. Future studies with larger datasets should incorporate uncertainty estimation, including confidence intervals and effect-size reporting, to strengthen statistical robustness and external validity.

FUNDING

This research was funded by the Scientific Research Coordination Unit of Anadolu University (Project No. SÇB-2024-2587). The funder had no role in study design, data collection, analysis, or publication decisions.

CONFLICT OF INTEREST STATEMENT

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and publication of this article.

REFERENCES

[1] S. Ahmad *et al.*, "Artificial intelligence and its role in education," *Sustainability*, vol. 13, no. 22, p. 12902, 2021, DOI: 10.3390/su132212902.

[2] L. Chen, P. Chen, and Z. Lin, "Artificial Intelligence in education: A review," *IEEE Access*, vol. 8, pp. 75264–75278, 2020, DOI: 10.1109/ACCESS.2020.2988510.

[3] S. Wang *et al.*, "Artificial intelligence in education: A systematic literature review," *Expert Syst. Appl.*, vol. 252, pp. 124167, 2024, DOI: 10.1016/j.eswa.2024.124167.

[4] K. Yamazaki, "Editorial: Future of technology in language teaching & learning," vol. 1, no. 1, pp. 1–2, 2018, DOI: 10.29140/TLTL.V1N1.153.

[5] M. Zafari *et al.*, "Artificial intelligence applications in K-12 education: A systematic literature review," *IEEE Access*, pp. 1–1, 2022, DOI: 10.1109/ACCESS.2022.3179356.

[6] H. Crompton *et al.*, "AI and English language teaching: Affordances and challenges," *Br. J. Educ. Technol.*, vol. 55, pp. 2503–2529, 2024, DOI: 10.1111/bjet.13460.

[7] N. Hockly, "Artificial intelligence in english language teaching: The good, the bad and the ugly," *RELC. J.*, vo. 54, pp. 445–451, 2023, DOI: 10.1177/00336882231168504.

[8] E. Kasneci *et al.*, "ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.*, vol. 103, p. 102274, 2023, DOI: 10.1016/j.lindif.2023.102274.

[9] Y. Luo, "Advancing language learning: The impact and challenges of computer-assisted language learning (CALL)," *Appl. Comput. Eng*, vol. 93, no. 1, pp. 83–88, 2024, DOI: 10.54254/2755-2721/93/20240924.

[10] A. Abd-Alrazaq *et al.*, "Large language models in medical education: Opportunities, challenges, and future directions," *JMIR Med. Educ.*, vol. 9, p. E48291, 2023, DOI: 10.2196/48291.

[11] B. Dong *et al.*, "Large language models in education: A systematic review," 2024 6th International Conference on Computer Science and Technologies in Education (CSTE), 131–134, 2024, DOI: 10.1109/CSTE62025.2024.00031.

[12] P. Pesch, "Potentials and challenges of large language models (LLMs) in the context of administrative decision-making," *Eur. J. Risk Regul.*, vol. 15, no. 1, pp. 76–95, 2025, DOI: 10.1017/err.2024.99.

[13] O. Fagbohun *et al.*, "Beyond traditional assessment: Exploring the impact of large language models on grading practices," *J. Artif. Intell. Mach. Learn. Data Sci.*, vol. 2, no. 1, pp. 1–8, 2024, DOI: 10.51219/jaimld/oluwole-fagbohun/19.

[14] H. Jiang, "Applications and research gaps of LLM-based English-as-a-foreign-language education," *J. Educ. Humanit. Soc. Sci.*, vol. 39, pp. 429–434, 2024, DOI: 10.54097/jkhv4m38.

[15] M. E. Butler-Pascoe, "The history of call," *Int. J. Comput.-Assist. Lang. Learn. Teach.*, vol. 1, no. 1, pp. 16–32, 2011, DOI: 10.4018/ijcallt.2011010102.

[16] N. Garrett, "Computer-assisted language learning trends and issues revisited: Integrating innovation," *Mod. Lang. J.* vol. 93, no. s1, pp. 719–740, 2009, DOI: 10.1111/j.1540-4781.2009.00969.x.

[17] D. R. Levine, "Computer-based analytic grading for German grammar instruction," *ACM Sigcue Outlook*, vol. 7, no. 3, p. 38, 1973a, DOI: 10.1145/963553.963556.

[18] D. R. Levine, *Computer-Based Analytic Grading for German Grammar Instruction*. Stanford, CA: Psychology and Education Series, 1973b, https://files.eric.ed.gov/fulltext/ED074787.pdf.

[19] R. Poloju, "Revolutionizing english language teaching: The impact of technology on language learning," vol. 12, no. 4, pp. 55–61, 2024. DOI: 10.69758/gimrj/2412ivvxiip0007

[20] M. Thomas and K. Yamazaki, *Computer-Assisted Language Learning. Education*. Oxford University, pp. 3–13, 2023, DOI: 10.1093/obo/9780199756810-0305.

[21] G. Urbaite, "The role of technology in modern language education," vol. 1, no. 1, pp. 3–10, 2024, DOI: 10.69760/w00r1v81.

[22] N. Ima and A. Jihad, "Computer-assisted language learning: The impact in language education," vol. 2, no. 1, pp. 36–42, 2024, DOI: 10.70184/smkf7m80.

[23] T. C. Bang, "Technology integration in english language education" (pp. 131–157). IGI Global, 2024, DOI: 10.4018/979-8-3693-3294-8.ch007.

[24] K. S. Puhachova, "Transforming education with artificial intelligence: Challenges, opportunities, and future directions," *Bull. Sci. Educ.*, vol. 11, no. 17, pp. 475–484, 2023, DOI: 10.52058/2786-6165-2023-11(17)-475-484.

[25] D. Shadiyeva, "Revolutionizing language teaching: Integrating modern pedagogical technologies into education systems," *Int. J. for Res Appl. Sci. Eng. Technol.*, vol. 12, no. 3, pp. 1325–1327, 2024, DOI: 10.22214/ijraset.2024.59119.

[26] C. A. Chapelle and E. Voss, "20 years of technology and language assessment in language learning & technology," *Lang. Learn. Technol.*, vol. 20, no. 2, pp. 116–28, 2016, DOI: 10.10125/44464.

[27] W. Y. Leong, Y. Z. Leong, and W. S. Leong, "Artificial intelligence in education," *IET Conf. Proc.*, vol. 2024, no. 22, pp. 183–184, 2025, DOI: 10.1049/icp.2024.4341.

[28] Z. Majkić and D. Vranješ, "The integration of artificial intelligence across educational levels: From primary school to university," In 10th International Scientific Conference Technics, Informatic, and Education (pp. 391–394). Proceedings TIE 2024. University of Kragujevac, Faculty of Technical Sciences, Čačak, 2024, DOI: 10.46793/tie24.391m.

[29] S. Mr. Mishra, "Revolutionizing education: The impact of AI-enhanced teaching strategies," *Int. J. for Res Appl. Sci. Eng. Technol.*, vol. 12, no. 9, pp. 9–32, 2024, DOI: 10.22214/ijraset.2024.64127.

[30] M. Parker *et al.*, "A large language model approach to educational survey feedback analysis," ArXiv, abs/2309.17447, 2023, DOI: 10.1007/s40593-024-00414-0.

[31] B. B. Turdaliyevna, "Using artificial intelligence technologies in language teaching," *Int. J. Lit. Lang.*, vol. 4, no. 11, pp. 35–39, 2024, DOI: 10.37547/ijll/volume04issue11-08.

[32] J. Young, "The rise of artificial intelligence in education," *Int. J. Innov Res Dev.*, vol. 13, no. 2, pp. 74–83, 2024, DOI: 10.24940/ijird/2024/v13/i2/feb24019.

[33] J. Algaraady and M. Mahyoob, "ChatGPT's capabilities in spotting and analyzing writing errors experienced by EFL learners," *Arab World Engl. J.*, no. 9, pp. 3–17, 2023, DOI: 10.24093/awej/call9.1.

[34] R. D'Souza, "Scope of Artificial Intelligence in education," International Journal of Emerging Knowledge Studies, vol. 3, no. 9, pp. 670–676, 2024, DOI: 10.70333/ijeks-03-09-035.

[35] W. Dai *et al.*, "Assessing the proficiency of large language models in automatic feedback generation: An evaluation study," 2024, DOI: 10.35542/osf.io/s7dvy.

[36] M. Kostic *et al.*, "LLMs in automated essay evaluation: A case study," In Proceedings of the AAAI Symposium Series vol. 3, no. 1, pp. 143–147, 2024, Association for the Advancement of Artificial Intelligence (AAAI). DOI: 10.1609/aaaiss.v3i1.31193.

[37] A. Kundu and D. Barbosa, "Are large language models good essay graders?" arXiv.Org, abs/2409.13120, 2024, DOI: 10.48550/arxiv.2409.13120.

[38] S. Marmoah *et al.*, "An integration of AI and traditional methodology in the education field in order to: Transform the trends," In 2024 4th International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 884–889). 2024 4th International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). IEEE, 2024, DOI: 10.1109/icacite60783.2024.10616679.

[39] F. Mercorio *et al.*, "Disce aut Deficere: Evaluating LLMs proficiency on the INVALSI Italian benchmark," 2024, DOI: 10.48550/arxiv.2406.17535.

[40] V. D. Lai *et al.*, "ChatGPT beyond english: Towards a comprehensive evaluation of large language models in multilingual learning (version 1)," *ArXiv.*, 2023, DOI: 10.48550/ARXIV.2304.05613.

[41] P. Vadlapati, "Multilingual prompting in LLMs: Investigating the accuracy and performance," *n.a. Sci. J. Res Eng. Manag.*, vol. 08, no. 12, pp. 1–7, 2024, DOI: 10.55041/ijsrem17694.

[42] U. Padó, Y. Eryilmaz and L. Kirschner, "Short-answer grading for German: Addressing the challenges," *Int. J. Artif. Intell. Educ.*, vol. 34, no. 4, pp. 1321–1352, 2023, DOI: 10.1007/s40593-023-00383-w.

[43] E. Latif *et al.*, "G-SciEdBERT: A contextualized LLM for science assessment tasks in German," arXiv.Org, abs/2402.06584, 2024, DOI: 10.48550/arxiv.2402.06584.

[44] F. Yavuz, Ö. Çelik, and G. Yavaş Çelik, "Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments," *Br. J. Educ. Technol.*, vol. 56, no. 1, pp. 150–166, 2024, DOI: 10.1111/bjet.13494.

[45] Y. Chen, H. Chen, and S. Su, "Fine-tuning large language models in education," In 2023 13th International Conference on Information Technology in Medicine and Education (ITME) (pp. 718–723). 2023 13th International Conference on Information Technology in Medicine and Education (ITME). IEEE, 2023, DOI: 10.1109/itme60234.2023.00148.

[46] B. Laufer, J. Kleinberg, and H. Heidari,. "Fine-tuning games: Bargaining and adaptation for general-purpose models," In Proceedings of the ACM Web Conference 2024 (pp. 66–76). WWW '24: The ACM Web Conference 2024. ACM, 2024, DOI: 10.1145/3589334.3645366.

[47] P. Wang *et al.*, "Making large language models better reasoners with alignment," ArXiv, abs/2309.02144, 2023, DOI: 10.48550/arXiv.2309.02144.

[48] L. Shen *et al.*, "Fast fine-tuning using curriculum domain adaptation," pp. 296–303, 2023, DOI: 10.1109/crv60082.2023.00045.

[49] R. Zhang *et al.*, "Unsupervised domain adaptation with adapter," arXiv: Computation and Language, 2021, https://arxiv.org/abs/2111.00667.

[50] E. Latif and X. Zhai, "Fine-tuning ChatGPT for automatic scoring," arXiv.Org, abs/2310.10072, 2023, DOI: 10.48550/arxiv.2310.10072.

[51] E. Latif and X. Zhai, "Fine-tuning ChatGPT for automatic scoring," *Comput. Educ. Artif. Intell.*, vol. 6, p. 100210, 2024, DOI: 10.1016/j.caeai.2024.100210.

[52] A. Sharma *et al.*, "A critical evaluation of AI feedback for aligning large language models," ArXiv, abs/2402.12366, 2024, DOI: 10.48550/arXiv.2402.12366.

[53] A. Vassar *et al.*, "Towards pedagogical LLMs with supervised fine tuning for computing education," ArXiv, abs/2411.01765, 2024, DOI: 10.48550/arXiv.2411.01765.

[54] M. H. Atlı, "Grammatical gender in German language and the acquisition of this system Almancada dilbilgisel cinsiyet sitemi ve bu sistemin yabancı dil olarak edinimi," *J. New Results Sci.*, vol. 12, no. 2, pp. 837–857, 2015, DOI: 10.14687/IJHS.V12I2.3266.

[55] E. Yücel and H. Yılmaz, "Die deutsche Grammatik: Eine Herausforderung für den türkischen DaF-Lerner," *Alman Dili ve Kültürü Araştırmaları Derg.*, vol. 5, no. 2, pp. 164–173, 2023, DOI: 10.55143/alkad.1357310.

[56] A. Riemenschneider *et al.*, "Linguistic complexity in teachers' assessment of German essays in high stakes testing," *Assess. Writ.*, vol. 50, p. 100561, 2021, DOI: 10.1016/j.asw.2021.100561.

[57] E. Yuzar, "Incorporating communicative competence in assessment and english language teaching in multilingual settings," vol. 2, pp. 8–13, 2020, DOI: 10.31849/reila.v2i1.3864.

[58] C. Hulme *et al.*, "LanguageScreen: The development, validation and standardization of an automated language assessment app," 2023, DOI: 10.31219/osf.io/gq6fh.

[59] X. Huang *et al.*, "Breakpoints in iterative development and interdisciplinary collaboration of AI-driven automated assessment," In 2024 21st International Conference on Information Technology Based Higher Education and Training (ITHET) (pp. 1–10). 2024 21st International Conference on Information Technology Based Higher Education and Training (ITHET). IEEE, 2024, DOI: 10.1109/ithet61869.2024.10837673.

[60] C. Hulme *et al.*, "LanguageScreen: The development, validation, and standardization of an automated language assessment app," *Lang. Speech Hear Serv. Sch.*, pp. 1–14, 2024, DOI: 10.1044/2024_lshss-24-00004.

[61] R. Muehlhoff and M. Henningsen, "Chatbots im Schulunterricht: Wir testen das Fobizz-Tool zur automatischen Bewertung von Hausaufgaben," 2024.

[62] W. Xie *et al.*, "Grade like a human: Rethinking automated assessment with large language models (version 1)," *ArXiv.*, 2024, DOI: 10.48550/ARXIV.2405.19694.

[63] L. Wright and N. Demeure, "Ranger21: A synergistic deep learning optimizer (version 2)," *ArXiv.*, 2021, DOI: 10.48550/ARXIV.2106.13731.

[64] A. Bapna, N. Arivazhagan, and O. Firat, "Simple, scalable adaptation for neural machine translation," *Empir. Methods Nat. Lang. Process.*, pp. 1538–1548, 2019, DOI: 10.18653/V1/D19-1165.

[65] B. Muralidharan *et al.*, "Knowledge AI: Fine-tuning NLP models for facilitating scientific knowledge extraction and understanding," 2024, DOI: 10.48550/arxiv.2408.04651.

[66] Y. Choi *et al.*, "Assessment modeling: Fundamental pre-training tasks for interactive educational systems (version 6)," *ArXiv.*, 2020, DOI: 10.48550/ARXIV.2002.05505.

[67] Z. Zhang *et al.*, "Balancing specialized and general skills in LLMs: The impact of modern tuning and data strategy (version 1)," *ArXiv.*, 2023, DOI: 10.48550/ARXIV.2310.04945.

[68] G. Vrbancic and V. Podgorelec, "Transfer learning with adaptive fine-tuning," *IEEE Access*, vol. 8, pp. 196197–196211, 2020, DOI: 10.1109/ACCESS.2020.3034343.

[69] L. Guan, "Weight prediction boosts the convergence of AdamW (version 2)," *ArXiv.*, DOI: 10.48550/ARXIV.2302.00195.

[70] R. Llugsi *et al.*, "Comparison between Adam, AdaMax and Adam W optimizers to implement a weather forecast based on neural networks for the Andean city of Quito," In 2021 IEEE Fifth Ecuador Technical Chapters Meeting (ETCM) (pp. 1–6). 2021 IEEE Fifth Ecuador Technical Chapters Meeting (ETCM). IEEE, 2021, DOI: 10.1109/etcm53643.2021.9590681.

[71] S. Xie and Z. Li, "Implicit bias of AdamW: ℓ_∞ norm constrained optimization (version 1)," *ArXiv*, 2024, DOI: 10.48550/ARXIV.2404.04454.

[72] A. Bewersdorff *et al.*, "Assessing student errors in experimentation using artificial intelligence and large language models: A comparative study with human raters," arXiv.Org, abs/2308.06088, 2023, DOI: 10.48550/arxiv.2308.06088.

[73] L. Gao and H. Ahmad, "A literature review of language assessment scale," *Asian Pendidikan*, vol. 4, no. 1, pp. 65–72, 2024, DOI: 10.53797/aspen.v4i1.7.2024.

[74] M. Li, Q. Gao, and T. Yu, "Kappa statistic considerations in evaluating inter-rater reliability between two raters: Which, when and context matters," *BMC Cancer*, vol. 23, no. 1, pp. 799–804, 2023, DOI: 10.1186/s12885-023-11325-z.

[75] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochem Med (Zagreb)*, vol. 22, no. 3, pp. 276–282, 2012, DOI: 10.11613/BM.2012.031.

[76] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)," *Geosci. Model Dev. Discuss.*, vol. 7, no. 1, pp. 1525–1534, 2014, DOI: 10.5194/GMDD-7-1525-2014.

[77] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not," *Geosci. Model Dev.*, vol. 15, no. 14, pp. 5481–5487, 2022, DOI: 10.5194/gmd-15-5481-2022.

[78] T. O. Hodson, "Root mean square error (RMSE) or mean absolute error (MAE): when to use them or not," *Geo. Mod. Dev.*, vol. 15, no. 7, pp. 5481–5487, 2022, DOI: 10.5194/gmd-2022-64.

[79] S. Faraby, A. Romadhony, and K. Adiwijaya, "Analysis of LLMs for educational question classification and generation," *Computers Educ. Artif. Intell.*, vol. 7, p. 1002698, 2024, DOI: 10.1016/j.caeai.2024.100298.

[80] U. Farooq, S. A. Malik, and A. S. Syed, "Revolutionizing higher education: The transformative impact of artificial intelligence on teaching, learning, and career preparation," *ShodhKosh: J. Vis. Perform. Arts*, vol. 5, no. 3, pp. 789–794, 2024, DOI: 10.29121/shodhkosh.v5.i3.2024.3239.

[81] K. Kavitha and V. P. Joshith, "The transformative trajectory of artificial intelligence in education: The two decades of bibliometric retrospect," *J. Educ. Technol. Syst.*, vol. 52, no. 3, pp. 376–405, 2024, DOI: 10.1177/00472395241231815.

[82] D. Kazimova *et al.*, "Transforming university education with AI: A systematic review of technologies, applications, and implications," *Int. J. Eng. Pedagogy (iJEP)*, vol. 15, no. 1, pp. 4–24, 2025, DOI: 10.3991/ijep.v15i1.50773.

[83] K. Mohamed *et al.*, "Hands-on analysis of using large language models for the auto evaluation of programming assignments," *Inf. Syst.*, vol. 128, p. 102473, 2024, DOI: 10.1016/j.is.2024.102473.

[84] W. Miller, "Adapting to AI: Reimagining the role of assessment professionals," *Intersection J. Intersection Assess. Learn.*, vol. 5, no. 4, pp. 99–113, 2024, DOI: 10.61669/001c.121439.

[85] B. Turnbull, "Towards new standards in foreign language assessment: Learning from bilingual education," *Int. J. Biling. Educ. Biling.*, vol. 23, pp. 488–498, 2020, DOI: 10.1080/13670050.2017.1375891.

[86] S. C. Vetrivel, P. Vidhyapriya, and V. P. Arun, "The role of AI in transforming assessment practices in education," in *Advances in Educational Marketing, Administration, and Leadership Book Series*, pp. 43–70, 2024, DOI: 10.4018/979-8-3693-5443-8.ch003.