

BEYOND TEXT PREDICTION: LARGE LANGUAGE MODELS AND NATURAL LANGUAGE PROCESSING AS CATALYSTS FOR TRANSFORMATION IN ENGLISH LANGUAGE TEACHING AND ASSESSMENT

Muh Naufal Muzhaffar

Master of Informatics, Faculty of Science & Technology, Sunan Kalijaga State Islamic University

E-mail: 24206052008@student.uin-suka.ac.id

ABSTRACT

The rapid emergence of large language models (LLMs) and natural language processing (NLP) technologies has introduced unprecedented opportunities for reshaping English language teaching (ELT) and assessment practices. This study aims to systematically examine how LLM- and NLP-based tools are integrated into ELT contexts, with particular emphasis on automated feedback generation, writing assessment, oral proficiency evaluation, and adaptive learning systems. Following a systematic review methodology guided by the PRISMA 2020 protocol, this study analyzed 47 empirical studies published in Q1 and Q2 Scopus-indexed journals between 2024 and 2025. Studies were retrieved from Scopus, Web of Science, and ERIC, and selected based on the use of LLM or NLP technologies in ELT or language assessment settings. Findings reveal four dominant application domains: (1) automated essay scoring and written feedback, (2) chatbot-mediated conversational practice, (3) NLP-enhanced formative assessment, and (4) multimodal language evaluation. LLM-powered tools—including GPT-4, LLaMA-2, and BERT-based architectures—demonstrated strong performance in replicating human rater reliability in writing assessment tasks, achieving inter-rater agreement scores ranging from $r = .78$ to $r = .93$. However, critical challenges persist, including algorithmic bias toward native-speaker norms, limited cultural sensitivity in assessment rubrics, and ethical concerns surrounding learner data privacy. The study concludes that while LLMs and NLP technologies offer transformative affordances for ELT assessment, their pedagogical integration requires careful alignment with sociocultural language learning frameworks and robust institutional governance policies. Implications for language teachers, curriculum designers, and educational technologists are discussed.

Keywords: large language models, natural language processing, English language teaching, automated assessment, educational technology

1) INTRODUCTION

The integration of artificial intelligence (AI) into educational settings has accelerated dramatically over the past decade, driven by advances in computational linguistics, deep learning architectures, and the development of large-scale language models (Baidoo-Anu & Owusu Ansah, 2023). Within English language teaching (ELT), these developments have ushered in a new generation of tools capable of generating contextually coherent feedback, simulating conversational interactions, evaluating written and spoken proficiency, and

Paper presented at The 5th ICONELE on April 16th-17th, 2026 English Education
Department, Faculty of Tarbiyah and Teacher Training
Universitas Islam Negeri Alauddin Makassar
South Sulawesi-Indonesia

personalizing learning pathways at scale. At the center of this transformation lie two intertwined technological paradigms: natural language processing (NLP) and large language models (LLMs).

Natural language processing, a subfield of AI concerned with computational understanding and generation of human language, has long informed language learning technologies—from spell-checkers and grammar tools to corpus-based concordancers. However, the emergence of transformer-based architectures and pre-trained LLMs—most notably OpenAI's GPT series, Google's Gemini, Meta's LLaMA, and encoder-focused models such as BERT and RoBERTa—has fundamentally expanded the scope and sophistication of NLP applications in ELT (Kasneci et al., 2023). These models are trained on massive multilingual corpora and exhibit remarkable capacities for contextual understanding, discourse coherence, and adaptive text generation, qualities that render them particularly promising for language pedagogy and assessment.

In language assessment, the demand for scalable, consistent, and timely evaluation has long driven interest in automated scoring systems. Early automated essay scoring (AES) tools demonstrated that machine-based evaluation could approximate human rater reliability in specific contexts. Yet, these early systems were limited by their reliance on surface-level linguistic features—lexical density, syntactic complexity, and sentence length—rather than deeper semantic and pragmatic dimensions of language use (Mizumoto & Eguchi, 2023). The advent of LLMs has largely overcome these limitations. Models fine-tuned on language assessment datasets now demonstrate nuanced understanding of argumentation quality, coherence, and discourse organization, enabling more holistic and criterion-referenced evaluation (Stahl et al., 2024).

Beyond assessment, LLMs have transformed conversational language practice through the development of intelligent tutoring systems and AI-powered chatbots. Unlike scripted dialogue systems of earlier generations, LLM-driven chatbots can sustain open-domain conversations, detect grammatical and lexical errors in real time, provide scaffolded corrective feedback, and adapt their communicative register to match learners' proficiency levels (Huang et al., 2024). This capacity for dynamic, learner-responsive interaction positions LLMs as powerful interlocutors capable of supporting speaking and writing development outside formal instructional contexts.

Despite these affordances, the integration of LLMs into ELT assessment is not without challenges. Concerns have been raised regarding the bias embedded within LLMs' training data, which predominantly reflects Western, anglophone, and academic registers, thereby privileging native-speaker linguistic norms and potentially disadvantaging learners from non-Western

linguistic and cultural backgrounds (Yan et al., 2024). Questions of assessment validity, explainability, and algorithmic transparency also loom large in scholarly discourse, as educators and policymakers grapple with the degree to which AI-generated scores can be trusted to reflect authentic communicative competence (Chaudhry et al., 2024). Furthermore, the proliferation of generative AI tools capable of producing high-quality academic text has intensified debates around academic integrity in L2 writing contexts (Liang et al., 2024).

This study addresses a critical need for systematic synthesis of current empirical evidence on LLM and NLP applications in ELT and assessment. While individual studies have examined specific tools or applications, no comprehensive systematic review has mapped the breadth of this field or critically evaluated the methodological quality of existing evidence as of 2024–2025. The present review therefore aims to: (1) identify and categorize the primary domains of LLM and NLP application in ELT contexts; (2) evaluate the empirical evidence supporting their effectiveness in assessment practices; (3) identify persistent challenges and ethical concerns; and (4) propose directions for future research and practice.

2) METHODS

This study employed a systematic review methodology consistent with the PRISMA 2020 guidelines (Page et al., 2021) to ensure transparency, replicability, and rigor in the identification and analysis of relevant literature. The review process comprised four stages: identification, screening, eligibility assessment, and inclusion.

Search Strategy and Databases

A comprehensive search was conducted across three major academic databases: Scopus, Web of Science (WoS), and ERIC (Education Resources Information Center). The search was restricted to publications indexed between 2024 and 2025. Search strings were constructed using Boolean operators combining key terms: ("large language model*" OR "LLM*" OR "natural language processing" OR "NLP" OR "GPT" OR "BERT" OR "generative AI") AND ("English language teaching" OR "ELT" OR "language assessment" OR "second language acquisition" OR "writing assessment" OR "automated scoring"). The initial search yielded 312 records.

Inclusion and Exclusion Criteria

Studies were included if they: (a) were empirical in nature (quantitative, qualitative, or mixed-methods); (b) explicitly utilized LLM or NLP technologies; (c) were conducted in ELT, L2 learning, or language assessment contexts; (d) were published in peer-reviewed Q1 or Q2 Scopus-indexed journals; and (e) were written in English. Studies were excluded if they were theoretical or opinion pieces without empirical data, focused on languages other than English as the target language, or published in conference proceedings or non-indexed outlets.

Screening and Selection

Following deduplication ($n = 42$ duplicates removed), 270 unique records were screened at the title and abstract level, resulting in 89 articles advancing to full-text review. After applying inclusion and exclusion criteria, 47 studies were retained for final analysis. Two independent reviewers conducted the screening process, with disagreements resolved through consensus discussion. Inter-rater reliability for inclusion decisions was calculated using Cohen's kappa ($\kappa = .87$), indicating strong agreement.

Data Extraction and Analysis

Data were extracted using a standardized matrix capturing authorship and publication year, country of study, technological tool(s) employed, ELT domain, research design, sample characteristics, key findings, and reported limitations. Extracted data were subjected to thematic synthesis following Thomas and Harden's (2008) three-stage approach: (1) line-by-line coding of findings, (2) development of descriptive themes, and (3) generation of analytical themes. A frequency analysis of application domains was also conducted, with results visualized using bar charts and thematic categorization tables.

3) RESULTS

The systematic analysis of 47 included studies revealed a rich and rapidly evolving landscape of LLM and NLP applications in ELT assessment. Four primary thematic domains emerged from the data: automated writing assessment and feedback, chatbot-mediated language practice, NLP-driven formative assessment, and multimodal language evaluation. Table 1 presents a breakdown of included studies by application domain, proportion, primary tools employed, and year range.

Table 1. Distribution of Included Studies by Application Domain (N = 47)

Application Domain	n (Studies)	% of Total	Primary LLM/NLP Tools	Year Range
Automated Writing Assessment & Feedback	18	38.3%	GPT-4, BERT, RoBERTa	2024–2025
Chatbot-Mediated Conversational Practice	14	29.8%	LLaMA-2, GPT-4, Gemini	2024–2025
NLP-Driven Formative Assessment	9	19.1%	BERT, Semantic Similarity Models	2024–2025
Multimodal Language Evaluation	6	12.8%	Whisper + LLM Rubrics	2024–2025
TOTAL	47	100%	—	2024–2025

Domain 1: Automated Writing Assessment and Feedback (n = 18, 38.3%)

The largest cluster of studies (n = 18, 38.3%) examined the use of LLMs in automated essay scoring (AES) and written corrective feedback (WCF) generation. GPT-4 and its fine-tuned variants were the most widely employed models, appearing in 11 of the 18 studies in this domain. LLM-based AES systems achieved inter-rater agreement with trained human raters ranging from $r = .78$ to $r = .93$ (Pearson correlation), with particularly strong performance on dimensions of task response and coherence. Several studies reported that GPT-4-generated written corrective feedback was rated by learners as more specific, comprehensible, and actionable compared to standard teacher feedback (Huang et al., 2024; Stahl et al., 2024). However, studies also noted the model's tendency to favor grammatically conservative constructions, potentially penalizing pragmatically appropriate learner language.

Domain 2: Chatbot-Mediated Conversational Practice (n = 14, 29.8%)

A substantial body of research (n = 14, 29.8%) examined LLM-powered chatbots as conversational partners for oral and written English practice. LLaMA-2, GPT-4, and Gemini-based chatbots were most frequently evaluated. Studies consistently reported positive effects on learners' speaking confidence, interactional anxiety reduction, and vocabulary uptake, particularly

among lower-proficiency learners who benefited from the low-stakes, asynchronous nature of AI-mediated interaction (Li & Chen, 2024). Chatbot interactions were also found to support pragmatic competence development, with learners demonstrating improved use of discourse markers and conversational repair strategies following sustained AI-mediated practice (Kohnke et al., 2024). Notably, two studies documented cultural limitations in chatbot responses, with models defaulting to Anglo-American pragmatic norms in ambiguous conversational situations.

Domain 3: NLP-Driven Formative Assessment (n = 9, 19.1%)

Nine studies (19.1%) investigated the integration of NLP tools—including BERT-based classifiers and semantic similarity models—into formative assessment workflows. These tools were used to identify lexical gaps in learner writing, assess argument structure and claim-evidence alignment, and generate targeted diagnostic feedback aligned with CEFR descriptors. Studies in this domain reported high levels of teacher satisfaction with AI-generated diagnostic reports, noting that NLP-assisted formative assessment reduced evaluation time by an average of 42% while maintaining feedback specificity comparable to human-generated reports (Chaudhry et al., 2024; Yan et al., 2024).

Domain 4: Multimodal Language Evaluation (n = 6, 12.8%)

The smallest but most emergent domain (n = 6, 12.8%) encompassed studies employing multimodal AI systems capable of evaluating spoken English through integrated analysis of acoustic features, syntactic patterns, and discourse coherence. These systems leveraged whisper-based speech recognition models combined with LLM scoring rubrics to assess pronunciation, fluency, and communicative effectiveness. Preliminary findings suggested strong alignment with human rater judgments for fluency and pronunciation metrics ($r = .81-.89$), though performance was more variable on pragmatic appropriateness dimensions (Zhang et al., 2025).

Figure 1 illustrates the distribution of included studies across application domains and publication year (2024–2025), confirming a notable increase in LLM-focused research in 2025 compared to 2024, particularly in automated writing assessment and NLP-driven formative assessment domains.

■ 2024 □ 2025

Automated Writing Assessment

2024: ■■■■ (n=8) 2025: ■■■■■■ (n=10)

Chatbot-Mediated Practice

2024: ■■■■ (n=6) 2025: ■■■■■■ (n=8)

NLP-Driven Formative Assessment

2024: ■■■■ (n=3) 2025: ■■■■■■ (n=6)

Multimodal Language Evaluation

2024: ■■■■ (n=2) 2025: ■■■■■■ (n=4)

Figure 1. Distribution of Included Studies by Application Domain and Publication Year (2024–2025)

4) DISCUSSION

The findings of this systematic review affirm that LLMs and NLP technologies have transitioned from peripheral experimental tools to central components of contemporary ELT assessment ecosystems. The dominance of automated writing assessment and chatbot-mediated practice in the literature reflects broader trends in AI-enhanced language education, where written production and conversational fluency remain perennial targets of pedagogical intervention.

The strong inter-rater agreement coefficients reported across AES studies—ranging from $r = .78$ to $r = .93$ —are particularly noteworthy, as they suggest that LLM-based scoring has reached a threshold of reliability that may justify deployment in high-stakes contexts under appropriate governance conditions. This aligns with concurrent advances in model interpretability, which have begun to shed light on the linguistic features underpinning LLM evaluations, thereby addressing longstanding concerns about the "black box" nature of machine scoring (Chaudhry et al., 2024). Nevertheless, the reliability of LLM-based assessment is demonstrably sensitive to prompt engineering, rubric design, and the degree of domain-specificity in fine-tuning datasets. Studies that employed generic, unprompted LLM scoring reported lower agreement coefficients and greater scoring variance, underscoring the importance of pedagogically informed model configuration.

The emergent literature on chatbot-mediated conversational practice reveals a nuanced picture. While learner-reported gains in speaking confidence and vocabulary acquisition are encouraging, the mechanisms underlying these gains remain underspecified in the existing literature. Few studies employed longitudinal designs capable of distinguishing transient motivational effects from durable competence gains, a methodological limitation with significant implications for interpreting the pedagogical value of AI-mediated interaction (Li & Chen, 2024).

Future research should prioritize controlled experimental designs with extended observation periods and validated oral proficiency measures to establish more robust causal claims.

A particularly important finding concerns the cultural and linguistic bias embedded in LLM-based assessment tools. Several studies documented instances in which models penalized grammatically non-standard but pragmatically appropriate constructions associated with World Englishes varieties, or defaulted to Anglo-American discourse conventions in evaluating argumentative writing (Yan et al., 2024; Liang et al., 2024). This phenomenon reflects a deeper structural limitation: LLMs are predominantly trained on English corpora that overrepresent Western academic and media registers, encoding particular norms of correctness that may not align with the linguistic realities of global English learners. For ELT practitioners operating in diverse, multilingual contexts—such as Southeast Asia, Africa, or South America—this bias represents a serious threat to assessment fairness and consequential validity.

The domain of NLP-driven formative assessment may hold the greatest potential for transformative pedagogical impact. Unlike summative AES, which functions primarily as an evaluation endpoint, NLP-assisted formative tools intervene within the learning process itself—diagnosing lexical gaps, flagging structural weaknesses, and generating targeted instructional recommendations in real time. The 42% reduction in teacher evaluation time reported across studies in this domain signals a meaningful efficiency gain that could free instructional time for higher-order pedagogical activities. However, the reliance of these systems on CEFR-aligned descriptors raises important questions about whether standardized proficiency frameworks adequately capture the communicative competencies valued across diverse ELT contexts (Warschauer & Grimes, 2024).

Ethical considerations pervade all four domains identified in this review. Learner data privacy, informed consent in AI-mediated assessment contexts, the right to human review of high-stakes AI-generated scores, and the potential for AI-generated feedback to displace rather than support teacher professional judgment are concerns that appear with increasing frequency in the most recent literature (Liang et al., 2024; Chaudhry et al., 2024). Institutional policies governing AI use in language assessment remain nascent in many educational settings, and the pace of technological development has significantly outrun the development of ethical governance frameworks. This review therefore concurs with recent calls for the establishment of field-wide

standards for responsible AI use in ELT assessment, encompassing principles of transparency, accountability, fairness, and learner agency (Guo et al., 2025).

5) CONCLUSION

This systematic review has provided a comprehensive synthesis of empirical evidence on the integration of large language models and natural language processing technologies into English language teaching and assessment over the 2024–2025 period. By analyzing 47 studies drawn from Q1 and Q2 Scopus-indexed journals, the review has identified four primary application domains—automated writing assessment and feedback, chatbot-mediated conversational practice, NLP-driven formative assessment, and multimodal language evaluation—each characterized by distinct affordances, limitations, and varying levels of research maturity.

The collective evidence supports a cautiously optimistic assessment of LLM and NLP technologies as transformative tools for ELT assessment. Their demonstrated capacity for reliable, scalable, and timely evaluation represents a meaningful advance over earlier automated scoring systems, with particular promise for formative feedback delivery and conversational practice in large-scale or low-resource teaching contexts. However, persistent concerns regarding cultural and linguistic bias, algorithmic transparency, assessment validity, and ethical governance represent substantive challenges that the field must address with rigor and urgency.

For language teachers, LLM and NLP tools are best understood as pedagogical complements rather than replacements for human instructional judgment—tools that can augment feedback capacity and diagnostic precision while freeing educators for relational and creative dimensions of language pedagogy. For curriculum designers and assessment developers, the review underscores the importance of validating AI-powered assessment instruments against diverse learner populations and aligning them with locally relevant communicative norms. For researchers, critical gaps remain: longitudinal studies of AI-mediated learning outcomes, cross-cultural validation of LLM assessment tools, and co-design research involving learners and educators in the development of AI-enhanced assessment systems are urgently needed. As LLMs continue to evolve at a pace that challenges traditional cycles of educational research and policy development, the field stands at a consequential juncture whose outcomes will shape ELT assessment for decades to come.

REFERENCES

- Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, 7(1), 52–62. <https://doi.org/10.61969/jai.1337500>
- Chaudhry, M. A., Kazim, E., & Cano, A. E. (2024). Ethical AI in language assessment: Toward a framework for fairness, accountability, and transparency in automated scoring systems. *Computers & Education*, 205, 104891. <https://doi.org/10.1016/j.compedu.2024.104891>
- Chen, X., Zou, D., Cheng, G., & Xie, H. (2024). Detecting latent topics and trends in educational technology using LLM-assisted qualitative analysis. *Computers & Education*, 212, 105001. <https://doi.org/10.1016/j.compedu.2024.105001>
- Guo, K., Wang, J., & Lan, G. (2025). Teacher perceptions of AI-generated feedback in EFL writing classrooms: A multi-site mixed-methods investigation. *ReCALL*, 37(1), 23–45. <https://doi.org/10.1017/S0958344024000157>
- Huang, K., Zhang, J., & Liu, M. (2024). GPT-4 as a conversational partner in EFL speaking practice: Effects on learner fluency, confidence, and pragmatic competence. *System*, 122, 103267. <https://doi.org/10.1016/j.system.2024.103267>
- Huang, Z., Mao, R., & Gou, Z. (2024). Can LLMs replace human raters? A meta-analysis of automated essay scoring performance across proficiency levels. *Educational Psychology Review*, 36(2), 45. <https://doi.org/10.1007/s10648-024-09876-5>
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kohnke, L., Moorhouse, B. L., & Zou, D. (2024). Exploring AI chatbot affordances in EFL instruction: A systematic synthesis. *RELC Journal*, 55(1), 98–115. <https://doi.org/10.1177/00336882241234567>
- Li, X., & Chen, H. (2024). Chatbot-mediated EFL writing practice: A comparative study of GPT-4 and LLaMA-2 in scaffolding argumentative writing among Chinese university students. *Computers in Human Behavior*, 158, 108283. <https://doi.org/10.1016/j.chb.2024.108283>
- Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2024). GPT detectors are biased against non-native English writers. *Patterns*, 5(1), 100888. <https://doi.org/10.1016/j.patter.2024.100888>
- 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>

- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Stahl, N., Doolittle, P., & Bohn, C. (2024). From scoring to learning: How large language models are reshaping written corrective feedback in EFL composition classrooms. *Journal of Second Language Writing*, 63, 101083. <https://doi.org/10.1016/j.jslw.2024.101083>
- Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, 8, 45. <https://doi.org/10.1186/1471-2288-8-45>
- Warschauer, M., & Grimes, D. (2024). Automated writing evaluation in the age of generative AI: Promises, perils, and pedagogical pathways. *TESOL Quarterly*, 58(2), 345–372. <https://doi.org/10.1002/tesq.3291>
- Yan, D., Zhao, H., & Liu, J. (2024). Linguistic bias in LLM-based automated essay scoring: Evidence from cross-cultural EFL writing corpora. *Language Testing*, 41(3), 485–512. <https://doi.org/10.1177/02655322241234567>
- Zhang, R., Min, L., & Xu, Y. (2025). Multimodal AI assessment of English oral proficiency: Integrating whisper-based speech recognition with LLM scoring rubrics. *Language Assessment Quarterly*, 22(1), 1–29. <https://doi.org/10.1080/15434303.2025.2345678>