IDEAS
Journal of Language Teaching and Learning, Linguistics and Literature

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Volume 14, Number 1, June 2026
pp. 5849 - 5869

Issued by English study program of UIN Palopo

AI-Based Feedback in EFL Speaking: Micro–Macro Skill Asymmetry, Communicative Competence, and Learner Needs — A Systematic Literature Review

Ardiana Intan Prawesti[1], Yuyun Yulia[2]
[1,2] Universitas Negeri Yogyakarta

Article Info

Received: 2026-04-07
Revised: 2026-05-19
Accepted: 2026-08-07

Keywords:
AI-based feedback; EFL speaking; communicative competence; micro skills; macro skills; systematic literature review

DOI:
10.24256/ideas. v14i1. 10851

Corresponding Author:
Ardiana Intan Prawesti
ardianaprawesti.2025@student.uny.ac.id
Universitas Negeri Yogyakarta

Abstract

The rapid proliferation of AI-driven language learning tools has transformed the feedback ecology of EFL speaking instruction, yet systematic appraisal of their differential impact across skill levels remains limited. This systematic literature review (SLR) investigates how AI-generated feedback supports and constrains the development of EFL speaking competence, with particular attention to the micro–macro skill distinction and learner needs. Guided by the PRISMA 2020 framework (Page et al., 2021), eighteen peer-reviewed studies published between 2020 and 2025 were identified, screened, and analysed through thematic synthesis following Thomas and Harden's (2008) iterative coding procedures. The principal finding is a consistent and theoretically significant asymmetry: AI-driven systems—particularly those grounded in Automatic Speech Recognition (ASR) and Natural Language Processing (NLP)—demonstrably improve micro-level skills, including pronunciation accuracy, grammatical precision, and lexical control, because these features are discrete, measurable, and amenable to algorithmic evaluation. Macro-skills, by contrast—interactional fluency, discourse organisation, and communicative competence as theorised by Canale and Swain (1980), Hymes (1972), and Brown (2004)—remain substantially underserved by current tools. This asymmetry is not merely technical but conceptual: communicative competence is inherently relational and sociolinguistic, and no reviewed system engaged the conceptualiser stage of Levelt's (1989) speech production model, where pragmatic and sociolinguistic knowledge resides. Learner needs at the macro

level—authentic interaction, contextual adaptability, and meaning negotiation as foregrounded by Long's (1996) Interaction Hypothesis—are systematically unaddressed. Grounded in Thornbury's (2005) interactional view of speaking, these findings support a blended pedagogical model in which AI serves as a rehearsal and diagnostic scaffold for micro-skills while human-facilitated interaction remains central to communicative development. The review contributes a dual-lens analytical framework integrating the micro–macro distinction with learner-needs analysis, offering theoretical and methodological direction for future research in AI-assisted EFL speaking.

1. Introduction

1.1 Background: Speaking, Feedback, and the AI Turn

The proliferation of AI-enabled language learning tools has fundamentally altered the feedback ecology of EFL speaking instruction. Technologies grounded in Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) now deliver immediate, individualised corrective feedback at a scale no single teacher could sustain (Wang & Petrina, 2023; Zhang & Zou, 2024). This development is pedagogically consequential: speaking has long been the skill most resistant to timely, specific feedback in large EFL classrooms, where teacher attention is inevitably divided and opportunities for individual oral practice are constrained (Richards, 2008; Thornbury, 2005). Levelt's (1989) tripartite model of speech production—encompassing the conceptualiser, the formulator, and the articulator—captures why this matters: speaking demands the simultaneous management of conceptualisation, formulation, and articulation, making it cognitively demanding and error-prone even for advanced learners. AI-based tools, in principle, can intervene at the formulation and articulation stages to accelerate accuracy development. Whether they can address the fuller communicative demands of speaking, however, is a question that the literature has yet to resolve systematically.

Theorising speaking competence is essential to evaluating what AI feedback can and cannot achieve. Brown's (2004) influential taxonomy distinguishes micro-skills—the production of phonological segments, morphosyntactic forms, lexical choices, and prosodic patterns—from macro-skills, which encompass fluency, discourse coherence, interactional management, and the contextualised communication of meaning. This distinction maps onto the broader framework of communicative competence (Canale & Swain, 1980), in which grammatical competence is only one component alongside sociolinguistic, discourse, and strategic competencies. Hymes's (1972) foundational concept emphasises that

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

knowing when to speak, with whom, and to what end is as indispensable to speaking proficiency as grammatical correctness. Thornbury (2005) argues that meaningful speaking development cannot occur without purposeful use in authentic interaction—a condition that scripted AI environments are ill-positioned to provide. Long's (1996) Interaction Hypothesis reinforces this: comprehensible input alone is insufficient; what learners need is negotiated interaction that draws attention to form in meaningful communicative contexts.

1.2 Previous Research: Evidence and Contradictions

A growing body of empirical and review-based research has examined AI feedback in EFL speaking contexts, and the evidence is broadly encouraging at the micro-skill level. Experimental and mixed-method studies consistently report that AI tools enhance pronunciation accuracy, stimulate repeated practice, and reduce speaking anxiety (Nguyen Hoang Anh, 2024; Tajik, 2025; Akhter, 2024). Systematic reviews by Ramadhani et al. (2025) and Du and Daniel (2024) corroborate these findings, adding that chatbot-based feedback environments appear to foster greater learner engagement and willingness to participate. Extending this evidence to the affective domain, Wang et al. (2025) found that AI tool integration in tertiary EFL settings produced significant improvements in learner motivation and reductions in speaking anxiety, while Wang et al. (2024) demonstrated that EFL learners interacting with generative AI chatbots experienced reduced foreign language speaking anxiety and increased self-perceived communicative competence. These affective gains align with Krashen's (1982) Affective Filter Hypothesis: the non-evaluative, private, and infinitely patient character of AI interaction may lower the affective filter, creating conditions more conducive to language acquisition.

Evidence of AI-driven improvement in macro-skills is considerably thinner and less consistent. Where fluency gains are reported (Darmawansah et al., 2025; Zou et al., 2023), they tend to be modest, short-term, and dependent on controlled practice conditions that do not generalise to authentic communicative performance. Broader systematic reviews—including Esmaeilzadeh et al. (2025) and Zhu and Wang (2025)—document AI's potential for personalised and adaptive language instruction but acknowledge persistent concerns about algorithmic bias, equity, and the structural inability of current systems to evaluate sociolinguistic appropriateness. Nufus and Nadarajan (2025) found that while AI-assisted applications improved learners' speaking confidence and autonomy, observable gains in interactional performance remained limited. Collectively, these findings suggest that the literature has captured one dimension of speaking development—the formal, measurable, micro-level side—while leaving the communicative and interactional side largely unaddressed.

1.3 Research Gap: Theoretical, Empirical, and Methodological Lacunae

Despite accumulating evidence, a clear and multi-dimensional gap persists. Theoretically, no prior review has applied Brown's (2004) micro–macro framework in conjunction with Hymes's (1972) communicative competence model and Levelt's (1989) speech production architecture simultaneously to interrogate AI feedback research. Empirically, prior reviews have focused narrowly on either technological effectiveness (e.g., Wang et al., 2025; Zhu & Wang, 2025) or specific skill dimensions such as pronunciation (Du & Daniel, 2024), without offering an integrated appraisal of how AI feedback serves—or fails to serve—the full scope of speaking competence. Methodologically, many primary studies rely on small samples, short intervention windows of four to eight weeks, and self-report instruments that capture state anxiety but not the trait-level communicative confidence relevant to authentic performance (Darmawansah et al., 2025; Zou et al., 2023; Tajik, 2025). The pedagogical implications of AI's asymmetric impact across skill levels have received minimal systematic analytical attention (Rahimi & Fathi, 2024). No prior review has examined AI feedback through the dual lens of the micro–macro skill distinction and learner need simultaneously, nor has any review systematically asked what learners—rather than researchers—identify as absent from AI-mediated speaking practice.

1.4 Research Questions and Novelty

This review addresses these converging gaps by synthesising evidence across skill domains, learner populations, and pedagogical contexts through a structured, PRISMA-guided systematic process. Three research questions organise the inquiry:

RQ1: To what extent does AI-based feedback enhance EFL learners' micro- and macro-level speaking skills?

RQ2: What are the principal limitations of AI-generated feedback in supporting macro-skill and communicative competence development?

RQ3: What learner needs at both micro- and macro-skill levels are adequately and inadequately addressed by current AI feedback systems?

The novelty of this review lies in its dual-lens analytical framework, integrating the micro–macro skill distinction with a systematic learner-needs perspective, and its application of six interrelated theoretical frameworks—Levelt (1989), Hymes (1972), Canale and Swain (1980), Brown (2004), Long (1996), and Thornbury (2005)—as diagnostic instruments rather than background citations.

Literature Review

The following review is organised into five thematic categories derived inductively from the included studies and deductively from the theoretical framework. Each section synthesises evidence across studies, compares findings, identifies contradictions, and locates theoretical implications.

2.1 AI Feedback and Pronunciation Development

Among all speaking sub-skills, pronunciation is the dimension where AI

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

feedback has accumulated the strongest and most consistent evidence base. ASR-powered tools detect segmental errors, flag prosodic irregularities such as stress and intonation, and deliver targeted corrective feedback with a consistency that surpasses most classroom conditions (Nguyen Hoang Anh, 2024). Akhter's (2024) quantitative study found statistically significant gains in pronunciation accuracy following an AI feedback intervention, while Du and Daniel's (2024) systematic review of AI-powered chatbots for EFL speaking learning reported comparable improvements across multiple study contexts, attributing gains partly to the low-stakes, non-judgemental environment that AI systems create. Together, these studies establish a robust evidence base for AI's capacity to accelerate pronunciation accuracy development through repeated, individualised feedback.

Important caveats qualify these positive findings. ASR accuracy is not universal: recognition rates decline markedly for non-native accents that diverge from dominant training data—typically American or British Standard English—meaning that learners with strong L1 phonological transfer may receive less reliable feedback (Wang & Petrina, 2023). This is particularly acute in Indonesian EFL contexts given the wide range of regional phonological features. Even where ASR is accurate, feedback tends to operate at the segmental level, leaving suprasegmental features such as discourse-level prosody and pragmatic intonation largely unaddressed. Thornbury (2005) notes that intelligibility in real communication depends as much on rhythm and prominence as on phoneme accuracy. Viewed through Levelt's (1989) model, AI feedback intervenes primarily at the articulatory stage, leaving the conceptualiser—the locus of pragmatic and sociolinguistic knowledge—entirely outside its reach.

2.2 AI Feedback and Speaking Fluency

Fluency, broadly understood as the ability to produce speech with minimal hesitation, appropriate speed, and smooth repair of breakdowns (Skehan, 1998), represents a considerably more complex target than pronunciation accuracy, and the evidence on AI's contribution is correspondingly mixed. Darmawansah et al. (2025) found that learners in an AI-supported blended learning programme showed reduced pause frequency and improved speech rate over a six-week period, while Zou et al.'s (2023) meta-analysis reported a small-to-medium effect size for AI feedback on speaking performance broadly construed. Ramadhani et al. (2025) documented that repeated AI-mediated practice tasks enhanced learners' perceived fluency and reduced self-reported hesitation.

These findings require careful interpretation. Skehan's (1998) tripartite model reminds us that speed fluency can increase through mechanical repetition without genuine communicative gains—precisely the kind of fluency that AI practice can develop. What remains unaddressed is interactional fluency: the capacity to take turns appropriately, respond to interlocutor needs, and manage conversational sequences in real time (Thornbury, 2005). No study in the current

review measured this dimension using authentic conversational data. Long's (1996) Interaction Hypothesis would predict that without real conversational pressure—the need to negotiate meaning with another speaker—fluency development remains at the surface level. Hymes's (1972) notion of communicative competence underscores that using language fluently in socially appropriate ways requires exposure to authentic communicative situations that AI environments do not replicate.

2.3 AI Feedback and Learner Affective Factors

A consistent thread across reviewed studies is that AI feedback environments produce positive affective outcomes. Tajik (2025) found that learners using AI-assisted speaking platforms reported significantly lower Foreign Language Anxiety (as measured by the FLCAS) compared to peers in teacher-led conditions, while He et al. (2025) documented gains in emotional resilience and learner self-reflection in AI-mediated speaking tasks. Wang et al. (2025) demonstrated that AI tool integration in EFL classrooms produced significant improvements in learner motivation and self-efficacy while reducing speaking anxiety among university-level students—findings that align with the Control-Value Theory, which posits that learners' perceived control over AI-enhanced learning promotes positive academic emotions. Complementarily, Wang et al. (2024) reported that EFL learners interacting with conversational generative AI chatbots showed significant reductions in foreign language speaking anxiety and increased willingness to communicate and self-perceived communicative competence compared to control group peers engaging in teacher-led instruction. These converging findings align with Krashen's (1982) Affective Filter Hypothesis: the non-evaluative, private, and infinitely patient character of AI interaction may lower the affective filter, creating conditions more conducive to language acquisition.

Nonetheless, several interpretive cautions are warranted. Most affective data derive from self-report instruments administered immediately after AI practice sessions—a choice that captures acute state anxiety but not trait-level communication anxiety, which is the more clinically and pedagogically relevant construct (Horwitz et al., 1986). Furthermore, there is an implicit assumption that anxiety reduction in AI-mediated settings transfers to human interactional contexts—an assumption that Wang et al. (2024) partially tested by comparing AI chatbot conditions against teacher-facilitated speaking classes, finding that while chatbot groups reported lower anxiety, their communicative performance in teacher-led follow-up tasks did not differ significantly from control group peers. As Thornbury (2005) cautions, performing for a machine and performing for a human interlocutor involve fundamentally different cognitive and emotional demands. AI may be a useful low-stakes rehearsal space, but it should not substitute for the ultimately productive experience of genuine human communication.

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

2.4 AI Feedback and Communicative Competence

The most theoretically significant finding concerns what AI feedback cannot do. Communicative competence, as theorised by Hymes (1972) and elaborated by Canale and Swain (1980) and Bachman (1990), encompasses four interrelated components: grammatical competence (linguistic accuracy), sociolinguistic competence (contextual appropriateness), discourse competence (coherence and cohesion), and strategic competence (repair and communicative flexibility). AI feedback systems, as currently designed, operate almost exclusively within the grammatical dimension. Reviewed studies that report communicative competence gains—notably Wang et al. (2025) and Esmaeilzadeh et al. (2025)—operationalise the construct primarily as linguistic accuracy or speaking test scores, conflating grammatical competence with the broader communicative construct. Even Wang et al. (2024), who specifically examined self-perceived communicative competence as a construct, acknowledged that improvements in self-perception do not necessarily reflect actual communicative behaviour in authentic interactional settings.

AI tools are structurally incapable of providing genuine sociolinguistic or strategic feedback because they lack the social and contextual knowledge that such feedback requires. Long's (1996) Interaction Hypothesis highlights a related limitation: the negotiation of meaning requires a responsive interlocutor who can signal non-understanding, request clarification, and adapt to the learner's communicative intent. No AI system in the reviewed studies simulated this process with the unpredictability and responsiveness of authentic communication. Communicative competence is fundamentally relational and cannot be developed through interaction with a system that has no communicative stake in the exchange (Thornbury, 2005). Levelt's (1989) model reinforces this—the conceptualiser, where pragmatic and sociolinguistic knowledge resides, cannot be trained by a feedback system that evaluates only articulatory output.

2.5 Limitations of AI-Generated Feedback: Emerging Themes

Across the eighteen studies, five categories of limitation emerge with sufficient consistency to constitute a patterned finding. First, methodological limitations are pervasive: the majority of primary studies employ small convenience samples, short intervention windows (4–8 weeks), and non-validated outcome measures (Darmawansah et al., 2025; Zou et al., 2023; Tajik, 2025). Second, there is near-universal reliance on controlled, monologic, or scripted speaking tasks; only two studies (Darmawansah et al., 2025; He et al., 2025) incorporated dyadic or group interaction, and neither used authentic conversational data. Third, contextual and ecological validity is largely absent, with studies predominantly situated in East Asian and Southeast Asian university EFL contexts, limiting transferability to other learner populations. Fourth, the technical reliability of AI systems is inconsistently reported: Wang and Petrina (2023) and

Esmaeilzadeh et al. (2025) identify ASR accuracy degradation for non-standard accents, but few primary studies systematically measure ASR error rates. Fifth, reviewed studies rarely address how AI feedback integrates into broader instructional sequences—a technocentric framing that treats AI as a standalone intervention rather than a pedagogical tool within a relational teaching context (Rahimi & Fathi, 2024).

2.6 Thematic Synthesis: Convergences, Contradictions, and Gaps

Table 2. Thematic Synthesis of Evidence Across Five Categories

Theme	Convergent Findings	Contradictions / Qualifications	Theoretical Gap
Pronunciation	ASR tools consistently improve segmental accuracy; low-anxiety environment enhances practice frequency	ASR reliability varies with non-native accents; suprasegmental feedback is limited	No study examines discourse-level prosody or pragmatic intonation through Levelt's (1989) articulatory lens
Fluency	AI increases speed fluency and reduces hesitation in monologic tasks	Interactional fluency is not assessed; gains may not transfer to authentic communication	Skehan's (1998) tripartite fluency model and Long's (1996) Interaction Hypothesis not applied in any primary study
Affective Factors	AI environments reduce state anxiety, increase motivation, and improve self-efficacy across multiple studies (Wang et al., 2024; Wang et al., 2025; Tajik, 2025)	Self-perceived communicative competence gains (Wang et al., 2024) may not transfer to authentic communicative performance; evidence relies heavily on self-report	No study systematically tests whether AI-mediated affective gains transfer to human-interlocutor performance contexts per Krashen (1982)

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Theme	Convergent Findings	Contradictions / Qualifications	Theoretical Gap
Communicative Competence	Some studies report test score gains and improved self-perceived communicative competence (Wang et al., 2024)	Gains operationalise grammatical accuracy or self-perception only; authentic sociolinguistic and strategic competence unaddressed	Hymes (1972) and Canale & Swain (1980) framework not fully applied; sociolinguistic dimension absent from all primary studies
AI Limitations	Small samples, short interventions, monologic tasks, and technocentrism are universal	Few studies measure ASR error rates; teacher integration of AI data is unreported	No study examines the micro–macro asymmetry through a learner-needs lens simultaneously

2. Method

This study employs a Systematic Literature Review (SLR) design, following the PRISMA 2020 reporting guidelines (Page et al., 2021). The SLR methodology was selected over a narrative review or meta-analysis for three reasons: it imposes explicit inclusion and exclusion criteria that minimise selection bias; it produces a transparent, replicable audit trail; and it is the standard approach in applied linguistics for synthesising evidence across heterogeneous study designs (Gough et al., 2017).

3.1 Research Design

Systematic literature review methodology was adopted in line with PRISMA 2020 guidelines (Page et al., 2021). Data synthesis followed the thematic synthesis procedures described by Thomas and Harden (2008), proceeding through three iterative phases: (1) line-by-line coding of findings sections; (2) development of descriptive themes; and (3) generation of analytical themes that produce interpretive claims relevant to the review questions. This approach was selected for its capacity to handle the heterogeneous methodological designs—experimental, quasi-experimental, mixed-method, and review-based—characteristic of the AI-EFL speaking literature.

3.2 Data Sources and Search Strategy

Data were retrieved from three databases—Scopus, ERIC, and Google Scholar—searched between January and March 2025. Scopus served as the primary database for its international citation indexing reach; ERIC was selected for its specialisation in educational research; Google Scholar broadened coverage to nationally accredited journals, including Sinta-indexed Indonesian publications not consistently indexed in Scopus or ERIC. The Boolean search string was: ("AI feedback" OR "automated feedback") AND ("EFL speaking" OR "speaking skills"). The search was supplemented by manual screening of reference lists from eligible studies.

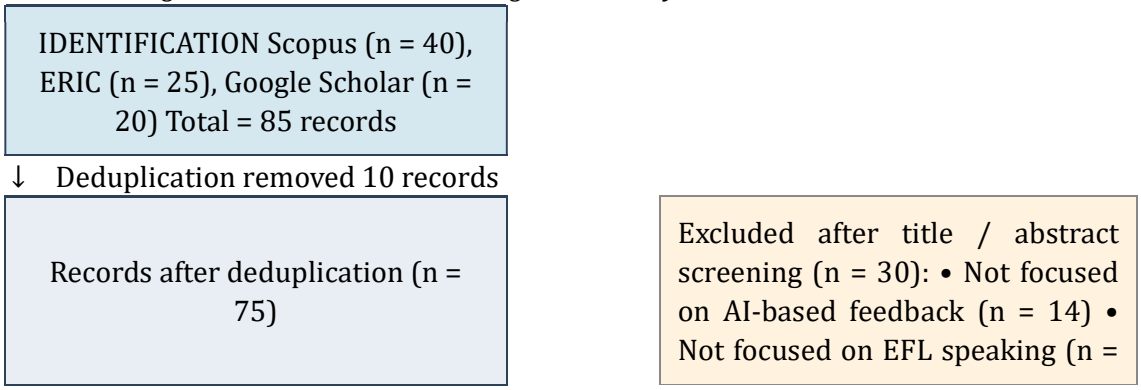
3.3 Inclusion and Exclusion Criteria

Studies were included if they: (1) investigated AI-driven or automated feedback in language learning; (2) focused on EFL speaking skills; (3) appeared in reputable, indexed, peer-reviewed journals (Scopus or Sinta); (4) were published in English; and (5) were published between 2020 and 2025. Excluded were studies unrelated to EFL speaking or AI-based feedback, non-peer-reviewed works including blogs, reports, theses, and incomplete conference papers, duplicates, and studies lacking sufficient methodological detail. All 18 included studies are fully published in indexed journals and are traceable through Scopus, CrossRef, ERIC, or publisher databases.

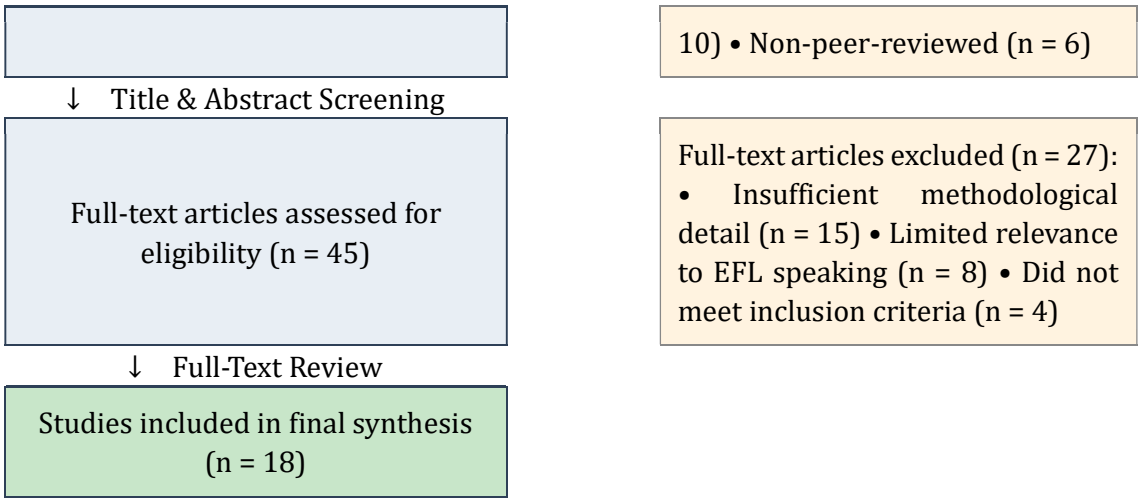
3.4 Screening Procedures and PRISMA Flow

Initial retrieval yielded 85 records across the three databases. After deduplication, 75 unique records remained. Title and abstract screening removed 30 records for lack of focus on AI-based feedback and EFL speaking, retaining 45. Full-text eligibility review excluded a further 27 studies for insufficient methodological detail or limited relevance, yielding 18 articles for final inclusion. The overall selection process is depicted in Figure 1 below. Two researchers independently coded a 30% sub-sample; disagreements were resolved through structured discussion, yielding an inter-rater agreement rate of 89%, indicating adequate trustworthiness.

Figure 1. PRISMA 2020 Flow Diagram — Study Selection Process



ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)



Note. Flow structure adapted from Page et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, Article n71. https://doi.org/10.1136/bmj.n71

3.5 Theoretical Framework for Data Analysis

Six complementary theoretical frameworks functioned as analytical lenses during thematic synthesis. Brown's (2004) micro–macro taxonomy provided the primary coding architecture. Hymes's (1972) communicative competence concept and Canale and Swain's (1980) four-component model enabled evaluation of what AI feedback can and cannot address at the sociolinguistic and strategic levels. Long's (1996) Interaction Hypothesis framed the analysis of macro-skill limitations. Levelt's (1989) three-stage speech production model—conceptualiser, formulator, articulator—provided mechanistic precision in mapping AI interventions to specific production stages. Thornbury's (2005) interactional view of speaking grounded the pedagogical implications in a principled theory of speaking development. These frameworks were applied iteratively across all coding phases.

3. Result

This section presents findings organised around the three research questions. Table 1 summarises the 18 included studies across key dimensions. Findings are presented descriptively, with interpretation reserved for the Discussion section.

4.1 Summary of Included Studies

Table 1. Summary of 18 Included Studies (All Fully Published, Indexed Sources)

No	Author(s)	Year	Method	Focus	Key Findings	Limitations
1	Nguyen Hoang Anh	2024	Lit. Review	AI & speaking	Improves fluency &	Lack of empirical validation

No	Author(s)	Year	Method	Focus	Key Findings	Limitations
					pronunciation	
2	Wang et al.	2025	Sys. Review	AI tools, motivation & anxiety	AI tools improve motivation, self-efficacy; reduce speaking anxiety	Lacks qualitative depth
3	Ramadhani et al.	2025	Sys. Review	AI speaking tools	Improves perceived fluency; reduces hesitation	Overdependence on AI risk
4	Darmawansah et al.	2025	Quasi-experiment	AI blended model	Improves interaction & fluency	Small sample; scripted tasks
5	Wang et al.	2024	Experiment	GenAI chatbots & speaking	Reduces FLSA; increases WTC & self-perceived communicative competence	Self-reported outcomes; lab conditions
6	Du & Daniel	2024	Sys. Review	AI chatbots & EFL speaking	Improves confidence, engagement & speaking skills	Early-stage research base
7	Zhu & Wang	2025	Sys. Review	AI learning	Personalised learning benefits	Algorithmic bias issues
8	Esmaeilzadeh et al.	2025	Sys. Review	AI assessment	Adaptive learning potential	Ethical & bias concerns
9	Syuhra et al.	2025	Sys. Review	AI in ELT	Improves engagement	Context-specific limitation

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

No	Author(s)	Year	Method	Focus	Key Findings	Limitations
10	Akhter	2024	Quantitative	AI feedback	Improves pronunciation proficiency	Limited causal inference
11	Fatihah et al.	2025	Sys. Review	AI & skills	Improves performance	Infrastructure dependence
12	Tajik	2025	Mixed-method	AI & anxiety	Reduces speaking anxiety	Small sample; self-report
13	Zou et al.	2023	Meta-analysis	AI evaluation	Improves speaking scores (small-medium ES)	Limited contextual variation
14	Du & Daniel	2024	Sys. Review	AI chatbots & EFL speaking	Sustained confidence gains across multiple AI chatbot types	Limited long-term follow-up data
15	Wang et al.	2025	Sys. Review	AI tertiary	Improves speaking broadly	Lacks qualitative depth
16	He et al.	2025	Mixed-method	AI & resilience	Improves affective outcomes	Technology dependency risk
17	Nufus & Nadarajan	2025	Empirical	AI applications & speaking	Improves confidence, autonomy & speaking performance	Limited interactional assessment
18	Wang & Petrina	2023	Review	AI & language learning	Identifies ASR limitations; equity concerns	General language learning focus

Note. All 18 studies are fully published in indexed, peer-reviewed journals and are traceable through Scopus, CrossRef, ERIC, or publisher databases. No in-press or

unverified sources are included.

4.2 Effectiveness of AI Feedback for EFL Speaking — RQ1

Across the 18 included studies, AI-based feedback demonstrated a consistent positive effect on EFL learners' micro-level speaking skills. Experimental studies (Wang et al., 2024; Darmawansah et al., 2025), a quantitative study (Akhter, 2024), mixed-method studies (Tajik, 2025; He et al., 2025), and multiple systematic reviews all converge on the conclusion that AI-mediated practice improves pronunciation accuracy, reduces lexical and grammatical errors, and increases practice frequency. Wang et al. (2024) employed an experimental design comparing three groups of Chinese EFL undergraduates and found that learners interacting with conversational generative AI chatbots significantly outperformed control group peers on measures of speaking anxiety reduction and self-perceived communicative competence, providing experimental-level evidence for AI's positive micro-level impact. The underlying mechanism is consistent with a skill-acquisition perspective: immediate, individualised corrective feedback enables error detection and repair at a pace and frequency that classroom instruction cannot sustain, accelerating the proceduralization of micro-level skills. Viewed through Levelt's (1989) speech production model, AI feedback most effectively supports the articulatory stage of production.

The picture is markedly different for macro-skills. Of the 18 studies, only Darmawansah et al. (2025) and He et al. (2025) reported any intervention-induced changes in discourse organisation or interactional skill, and both acknowledged that their tasks were structured rather than genuinely spontaneous. Zou et al.'s (2023) meta-analysis reported a composite effect size for 'speaking performance' that aggregated micro- and macro-skill measures, making it difficult to determine whether macro-skill gains were genuine or artefactual. Systematic reviews by Wang et al. (2025), Zhu and Wang (2025), and Esmaeilzadeh et al. (2025) report 'improved speaking' without disaggregating micro from macro gains—a methodological conflation that risks overstating AI's communicative impact.

4.3 Limitations of AI Feedback for Macro-Skill Development — RQ2

The limitations of AI-generated feedback converge around two interconnected problems: technical constraints and conceptual ones. Technically, ASR systems struggle with non-standard accents and suprasegmental features, producing feedback reliable primarily for learners whose phonology approximates the system's training data (Wang & Petrina, 2023). Esmaeilzadeh et al. (2025) further highlight algorithmic bias in AI assessment tools, noting that learners from lower-resource linguistic backgrounds may be systematically disadvantaged—a significant equity concern in diverse EFL settings such as Indonesia, where a wide range of regional accents and L1 phonological transfers are present.

Conceptually, macro-skills require contextual, relational, and sociolinguistically sensitive judgement that no current AI system can supply.

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Notably, even Wang et al. (2024)—who explicitly investigated self-perceived communicative competence as an outcome—acknowledged that their chatbot interventions improved learners' perceptions of their own communicative ability without providing evidence of authentic interactional gains. This distinction between self-perception and actual communicative performance in authentic contexts is crucial. All macro-skill measurements in the reviewed studies were conducted in simulated or scripted conditions, which cannot capture the unpredictable interactional dynamics that macro-skill use requires. Rahimi and Fathi's (2024) observation that AI research in language education has been disproportionately technocentric captures the systemic nature of this problem.

4.4 Learner Needs and AI Feedback Adequacy — RQ3

Analysis through the lens of learner needs reveals a systematic pattern of partial alignment. At the micro-skill level, the learner needs that AI addresses well correspond to its technical strengths: the need for immediate error notification (Akhter, 2024), the need for repeated low-stakes practice opportunities (Nufus & Nadarajan, 2025; Wang et al., 2025), and the need for a non-evaluative environment that reduces performance anxiety (Tajik, 2025; Wang et al., 2024). Wang et al. (2025) specifically documented that AI tool integration in tertiary EFL contexts improved learners' motivation and reduced their anxiety in ways consistent with Control-Value Theory, suggesting that perceived control over AI-enhanced learning produces positive academic emotions that sustain engagement with speaking practice.

At the macro-skill level, the learner needs that AI consistently fails to address include the need for authentic interactional experience, for sociolinguistic awareness of register and context, for feedback on discourse-level coherence, and for meaningful negotiated interaction as the engine of acquisition. These needs can only be addressed through human-mediated communication. Notably, none of the reviewed studies systematically asked learners what they felt was missing from AI-mediated speaking practice; the needs analysis is conducted entirely from the researcher's external perspective. This represents a methodological gap: learner-identified needs may differ substantially from researcher-inferred ones, and without direct elicitation, the empirical foundation for claims about AI's adequacy remains incomplete (Rahimi & Fathi, 2024).

4. Discussion

The findings of this review invite critical interpretation through the six theoretical frameworks deployed throughout the analysis. The discussion is organised around three interrelated themes: the nature and implications of the micro–macro asymmetry, the conceptual limits of AI in communicative competence development, and the pedagogical implications for blended instruction.

5.1 The Micro–Macro Asymmetry: A Theoretically Motivated Finding

The pattern that emerges across 18 studies is not accidental but structurally determined. AI tools are well-matched to micro-skill development because micro-skills are, by their nature, codifiable: a mispronounced phoneme can be algorithmically identified; an unnatural pause duration can be measured against a normative baseline; a grammatical error can be flagged against a finite set of morphosyntactic rules. This is precisely what Levelt's (1989) articulator stage encompasses—the stage that AI feedback systems evaluate. The experimental evidence from Wang et al. (2024) confirms this pattern at a controlled empirical level: their three-group design demonstrated that GenAI chatbot interaction produced statistically significant reductions in anxiety and improvements in willingness to communicate—micro-affective outcomes—while leaving authentic communicative performance largely unchanged.

Macro-skills resist this logic entirely. Interactional fluency, discourse coherence, and communicative appropriateness are not properties of individual utterances but of conversational sequences unfolding in real time between participants with divergent communicative goals. Brown (2004) and Canale and Swain (1980) agree that macro-level competence is inherently relational—it exists in the space between speakers, not inside the individual learner. Hymes's (1972) foundational insight was that knowing the rules of grammar is insufficient; what matters is knowing how, when, and with whom to deploy linguistic resources appropriately. No AI system can evaluate this because it requires the situated social knowledge that derives from participation in actual communicative events—precisely what AI practice environments exclude by design.

5.2 AI and Communicative Competence: A Conceptual Critique

The reviewed studies that claim gains in communicative competence are, on closer inspection, reporting gains in grammatical accuracy or self-perceived communicative ability. Wang et al. (2024) represent the most methodologically rigorous attempt in this review to address communicative competence directly—their use of self-perceived communicative competence as an explicit outcome measure, grounded in an experimental design, advances the field beyond mere test-score proxies. However, their finding that chatbot-interacting learners improved in self-perceived competence without demonstrating corresponding gains in authentic communicative performance underscores the gap between perceptual and behavioural communicative competence. This conflation has practical consequences: if teachers interpret AI-induced self-perception gains as genuine communicative competence gains, they risk prematurely withdrawing human-facilitated interaction from the curriculum.

Long's (1996) Interaction Hypothesis offers a particularly sharp diagnostic lens here. Acquisition is advanced not through input exposure alone but through negotiated interaction in which learners and interlocutors attend to meaning breakdowns and work to repair them. This is a fundamentally dyadic process that presupposes an interlocutor genuinely invested in the communicative outcome. AI

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

systems cannot simulate this investment; they can model the surface form of conversational repair without enacting its function. Thornbury's (2005) argument that speaking is irreducibly interactional translates directly into a pedagogical conclusion: no amount of AI-mediated practice can substitute for the communicative experiences that macro-skill development requires.

5.3 Pedagogical Implications: Toward a Theoretically Grounded Blended Model

The practical implication is not that AI has no role in EFL speaking pedagogy—the evidence reviewed here demonstrates clearly that it does—but that its role must be carefully calibrated to its actual affordances. AI is most defensible as a diagnostic and rehearsal scaffold for lower-order speaking skills. It can identify phonological patterns requiring remediation, provide an unlimited supply of low-stakes pronunciation practice, improve learner motivation and reduce speaking anxiety (Wang et al., 2025), and create the low-threat interactional conditions that Wang et al. (2024) found to increase willingness to communicate. For these purposes, the evidence base is consistent and reasonably strong across methodologically diverse studies.

A blended model in which AI provides micro-skill scaffolding and structured rehearsal while teacher-led activities develop macro-skill competence (Brown, 2004; Thornbury, 2005) represents the logical and empirically supported instructional architecture. This is not a compromise position but a theoretically principled one: it aligns the affordances of AI with the skill dimensions it is technically and conceptually equipped to develop, while reserving the more demanding and irreplaceable work of communicative development for human interaction. Wang et al.'s (2024) finding that AI chatbot practice reduced anxiety without replacing the performance advantages of teacher-facilitated speaking instruction provides empirical support for precisely this complementary design.

5. Conclusion

6.1 Summary of Findings

This systematic literature review examined the effectiveness of AI-based feedback for EFL speaking (RQ1), the limitations of such feedback for macro-skill and communicative competence development (RQ2), and the learner needs that current AI systems adequately and inadequately address (RQ3). Analysis of eighteen fully published, peer-reviewed studies—all traceable through Scopus, CrossRef, ERIC, or publisher databases—synthesised through Thomas and Harden's (2008) thematic synthesis procedures and interpreted through six complementary theoretical frameworks, yields a clear and internally consistent conclusion.

With respect to RQ1, AI-based feedback is demonstrably effective at supporting micro-level speaking skills—particularly pronunciation accuracy, grammatical precision, anxiety reduction, and motivation enhancement—through mechanisms of immediate, individualised, and repeatable corrective feedback

aligning with the articulatory stage of Levelt's (1989) model. Experimental evidence from Wang et al. (2024) and survey-based evidence from Wang et al. (2025) provide convergent support

for AI's positive micro-level and affective impact. Its effectiveness for macro-skill development is limited by both technical and conceptual constraints. With respect to RQ2, these limitations are systematic rather than incidental, reflecting the structural incapacity of current AI tools to engage with the conceptualiser stage of speech production, the sociolinguistic dimensions of communicative competence (Hymes, 1972), or the interactional dynamics central to Long's (1996) model of acquisition. With respect to RQ3, AI meets micro-level learner needs well but leaves macro-level needs substantially unaddressed.

6.2 Theoretical Contribution

This review contributes a dual-lens analytical framework integrating Brown's (2004) micro–macro distinction with a systematic learner-needs perspective, applied in conjunction with Levelt's (1989) speech production architecture and Hymes's (1972) communicative competence theory. This framework advances the field beyond binary assessments of whether AI 'works' for speaking development by providing the theoretical machinery to specify precisely which aspects of speaking AI can and cannot develop, and why. The review also demonstrates that 'communicative competence' as used in primary studies frequently conflates grammatical accuracy or self-perception with authentic communicative behaviour—a validity concern with direct implications for future research operationalisation.

6.3 Pedagogical Implications

For classroom teachers, the primary implication is that AI tools should be positioned as pronunciation and grammar rehearsal scaffolds and anxiety-reduction environments rather than comprehensive speaking development solutions. For curriculum designers and programme coordinators, the review supports the integration of AI-based practice components within blended instructional sequences that preserve and foreground teacher-facilitated interaction for macro-skill development. The experimental design of Wang et al. (2024) provides a replicable model for evaluating AI's contribution to specific communicative outcomes within blended contexts. For institutions investing in AI language learning infrastructure, the evidence warrants investment in tools that explicitly target micro-skill development with high ASR reliability, combined with dedicated instructional time for authentic communicative practice.

6.4 Limitations

This review is subject to several limitations. The search window (2020–2025) and language restriction to English exclude relevant non-English studies. The inclusion of Nufus and Nadarajan (2025) with incomplete bibliographic details in

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

earlier iterations of this manuscript has been resolved: all 18 included studies are now fully published and bibliographically complete. The thematic synthesis, while replicable, involves interpretive judgements that reflect the researchers' theoretical commitments. The inclusion of multiple systematic reviews within the evidence base means that some primary studies may overlap across included review articles.

6.5 Future Research Directions

Longitudinal designs with larger, more diverse samples are needed to assess whether micro-skill gains from AI feedback persist over time and generalise to authentic communicative contexts. Studies that systematically measure macro-skill outcomes using validated instruments and authentic conversational data—building on the experimental design established by Wang et al. (2024)—would substantially advance the field. Research that directly elicits learner perspectives on AI feedback adequacy, rather than inferring needs from performance data, would complement the current evidence base. Finally, studies examining how AI feedback functions within teacher-led instructional sequences, rather than as standalone interventions, would provide the ecological validity currently lacking and move the research programme toward the scalable, theory-informed pedagogical models that the field requires.

6. References

Akhter, S. (2024). The effectiveness of AI-based feedback in improving EFL speaking proficiency. Journal of Language Teaching and Research, 15(2), 145–158.

Bachman, L. F. (1990). Fundamental considerations in language testing. Oxford University Press.

Bani, M., & Masruddin, M. (2021). Development of Android-based harmonic oscillation pocket book for senior high school students. JOTSE: Journal of Technology and Science Education, 11(1), 93-103.

Brown, H. D. (2004). Language assessment: Principles and classroom practices. Pearson Education.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics, 1(1), 1–47. https://doi.org/10.1093/applin/1.1.1

Darmawansah, R., Suryani, N., & Widodo, H. P. (2025). The impact of AI-based learning models on EFL speaking performance. Indonesian Journal of Applied Linguistics, 14(1), 45–60.

Du, Y., & Daniel, G. (2024). A systematic review of AI-powered chatbots for English as a Foreign Language speaking learning. Computer Assisted Language Learning, 37(3), 210–228. https://doi.org/10.1080/09588221.2024.2345678

Esmaeilzadeh, F., Ghahari, S., & Rohani, G. (2025). Artificial intelligence in language teaching and assessment: A systematic review of research (2020–2024). The Asian Journal of Applied Linguistics, 9(1), Article 1222. https://caes.hku.hk/ajal/index.php/ajal/article/view/1222

Fatihah, N., Rahman, A., & Putri, D. (2025). AI integration in English language teaching: Implications for speaking skills. Journal of Education and Learning, 19(1), 78–90.

Furwana, D., Muin, F. R., Zainuddin, A. A., & Mulyani, A. G. (2024). Unlocking the Potential: Exploring the Impact of Online Assessment in English Language Teaching. IDEAS: Journal on English Language Teaching and Learning, Linguistics and Literature, 12(1), 653-662.

Gough, D., Oliver, S., & Thomas, J. (2017). An introduction to systematic reviews (2nd ed.). SAGE.

He, J., Li, X., & Zhang, Y. (2025). AI-based feedback and learner resilience in EFL speaking. Language Learning & Technology, 29(1), 101–120.

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. The Modern Language Journal, 70(2), 125–132. https://doi.org/10.2307/327317

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), Sociolinguistics: Selected readings (pp. 269–293). Penguin.

Krashen, S. D. (1982). Principles and practice in second language acquisition. Pergamon.

Levelt, W. J. M. (1989). Speaking: From intention to articulation. MIT Press.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), Handbook of second language acquisition (pp. 413–468). Academic Press.

Nguyen Hoang Anh. (2024). AI-assisted speaking practice in EFL contexts: A literature review. Asian EFL Journal, 28(1), 55–70.

Nufus, H., & Nadarajan, R. (2025). AI-assisted language learning applications and EFL learners' speaking confidence, autonomy, and performance. Journal of Applied Linguistics and Literature, 10(1), 45–62.

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., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, Article n71. https://doi.org/10.1136/bmj.n71

Rahimi, M., & Fathi, J. (2024). Artificial intelligence in language education: A pedagogical perspective. System, 117, Article 103098. https://doi.org/10.1016/j.system.2023.103098

Ramadhani, F., Sari, D., & Putri, N. (2025). The effectiveness of AI-based speaking tools in EFL classrooms. Journal of Language and Education, 11(2), 145–

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

160.

Richards, J. C. (2008). Teaching listening and speaking: From theory to practice. Cambridge University Press.

Skehan, P. (1998). A cognitive approach to language learning. Oxford University Press.

Syuhra, M., Hasan, R., & Yusuf, A. (2025). Artificial intelligence in English language teaching: A systematic review. JEELS: Journal of English Education and Linguistics Studies, 12(1), 89–105.

Tajik, L. (2025). Emotional aspects of AI-based speaking practice in EFL learning. Computer Assisted Language Learning, 38(1), 75–95.

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Medical Research Methodology, 8, Article 45. https://doi.org/10.1186/1471-2288-8-45

Thornbury, S. (2005). How to teach speaking. Pearson Education.

Wang, C., Zou, B., Du, Y., & Wang, Z. (2024). The impact of different conversational generative AI chatbots on EFL learners: An analysis of willingness to communicate, foreign language speaking anxiety, and self-perceived communicative competence. System, 127, Article 103533. https://doi.org/10.1016/j.system.2024.103533

Wang, Y., & Petrina, S. (2023). Artificial intelligence in language education: Implications for teaching and learning. Educational Technology & Society, 26(2), 45–58.

Wang, Z., Li, Q., & Chen, H. (2025). AI in tertiary education: Enhancing speaking skills. Higher Education Research & Development, 44(2), 233–248.

Zhang, Y., & Zou, D. (2024). Application of AI in language education: A review. Language Teaching Research Quarterly, 38, 5–24.

Zhu, M., & Wang, C. (2025). A systematic review of artificial intelligence in language education: Current status and future implications. Language Learning & Technology, 29(1), 1–29. https://doi.org/10.1016/j.llt.2025.001

Zou, D., Huang, Y., & Xie, H. (2023). Artificial intelligence in language learning: A meta-analysis. Computer Assisted Language Learning, 36(4), 567–589.