ChatGPT in Undergraduate English Language Majors: Benefits and Challenges for Writing and Speaking Proficiency

Nguyen Song Thao Anh[1]*, Nguyen Thi Kim Ngoc[1], Tran Khanh Bang[1], Huynh Thi Kim Ngan[1], Phan Nguyen Yen Quyen[1], Nguyen Thi Phuong Hong[2]

[1]Student of the High-Quality English Language Program, Cohort 48, School of Foreign Languages, Can Tho University, Vietnam

[2]School of Foreign Languages, Can Tho University, Vietnam

*** Corresponding Author**

DOI: https://dx.doi.org/10.47772/IJRISS.2025.910000364

Received: 20 October 2025; Accepted: 27 October 2025; Published: 12 November 2025

ABSTRACT

Large language models such as ChatGPT are reshaping undergraduate English language education. This systematic review synthesizes benefits and challenges for developing writing and speaking proficiency among English language majors. Following PRISMA-2020/PRISMA-S and SPAR-4-SLR, we searched Scopus, Web of Science, ERIC, EBSCO and SSRN, and analyzed 31 of the 708 empirical and review papers published. Findings converge on writing gains when ChatGPT is embedded in scaffolded, human-in-the-loop workflows that emphasize pre-writing, drafting, and revision. Effects are strongest for accuracy, coherence, and argument quality, and when teacher or peer moderation, prompt scaffolds, and transparent rubrics are present. For speaking, learners benefit through low-stakes practice, rehearsal and anxiety reduction; however, robust measurement lags behind, with scarce CEFR-aligned rubrics, limited voice-mode instrumentation, and few validated acoustic indicators. Integrity and equity remain central. Text-only AI detectors are brittle and sometimes unfair to non-native writers. Institutions should pivot to process-anchored assessment - combining prompt logs, version histories, and brief viva voce - to evidence authorship while preserving learning value. A forward research agenda is proposed, including a minimum reporting toolkit for speaking measurement (CEFR, ASR features, and ICC/κ) and multi-site randomized trials.

Keywords: assessment integrity, ChatGPT, human-in-the-loop pedagogy, L2 writing, oral proficiency

INTRODUCTION

Large language models (LLMs) such as ChatGPT are accelerating their penetration into higher education, offering scalable feedback, ubiquitous practice, and personalized scaffolding for undergraduate English language majors. The applied linguistics literature describes this domain as emergent yet rapidly developing: a systematic review synthesizing 70 empirical studies in English as a Second/Foreign Language (ESL/EFL) appeared within 18 months of ChatGPT's release, mapping applications across writing, speaking, assessment, and policy (Lo et al., 2024). Status-of-the-field essays consistently emphasize the nascency of the domain and call for clearer links between mechanisms and outcomes (Fang & Han, 2025; Han, 2024).

Writing: Classroom and blended-methods studies indicate that ChatGPT can supplement teachers by providing formative feedback, aiding ideation, and revision structuring, particularly in human-in-the-loop workflows (Teng, 2024). In undergraduate writing assessment, ChatGPT's scoring and qualitative comments show moderate-to-good consistency with human rater evaluation while scaling up the volume and type of feedback (Lu et al., 2024). Learner-centric studies report positive perceptions when ChatGPT is positioned as a "companion," with improvements in accuracy and coherence during drafting/revising (Tram et al., 2024). However, these benefits are uneven: without prompt scaffolds and rubric-based guidance, students may over-rely on surface-level edits, with limited transfer to higher-order reasoning skills (Han, 2024; Lo et al., 2024).

Speaking: The (emergent) voice and chat modes expand opportunities for low-stakes practice and can reduce anxiety (Tram et al., 2024). However, robust measurement of speaking proficiency gains - phonology/intonation, interactional competence - lags behind writing metrics. Validated CEFR-aligned rubrics and instrumentation (e.g., Automatic Speech Recognition (ASR) features for speech rate or pause ratio) remain underdeveloped (Fang & Han, 2025; Lo et al., 2024).

Challenges:

- Integrity and Equity: Academic integrity and the reliability/fairness of AI detection tools remain unresolved. Emerging evidence suggests that common detectors systematically misclassify non-native writers' (NNS) text as AI-generated (Liang et al., 2023).

- LLM Quality and Safety: LLMs can "hallucinate" references, homogenize writing styles, and produce feedback of variable specificity, reinforcing the need for teacher moderation and transparent AI usage policies (Lo et al., 2024; Teng, 2024).

- Methodological Gaps: Methodological gaps limit strong inference: many studies are short-term, single-site, or quasi-experimental, with sparse process data (e.g., keystroke logs, prompt histories) that could illuminate learning mechanisms (Fang & Han, 2025; Han, 2024).

Review Framework and Research Agenda: Framing the review within the TCCM Framework (Theory - Context - Characteristics - Methodology) clarifies contributions and gaps (Han, 2024). Contextually, the evidence is concentrated in Asian higher education and first-year writing, with fewer multi-site trials and speaking interventions (Fang & Han, 2025; Lo et al., 2024). At the characteristic level, task type, modality, human coordination, and initial proficiency are likely to moderate effects. Methodologically, longer interventions (~ 12 - 16 weeks), multi-institutional Randomized Controlled Trials (RCTs), and open reporting standards are priorities (Fang & Han, 2025; Han, 2024).

Against this background, our review synthesizes the benefits and challenges of ChatGPT for writing and speaking in undergraduate English programs and outlines a forward research agenda. We ask:

What measurable advantages and risks emerge when ChatGPT is integrated into writing and speaking curricula?

Under what pedagogical and assessment conditions are effects enhanced or mitigated?

What design principles and research infrastructure - validated rubrics, process analytics, pre-registration - are necessary to mature the evidence base?

METHODOLOGY

Design and reporting standards

We conducted a systematic review aligned with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA-2020) guidelines. To enhance review craft and transparency, we also followed the Scientific Procedures and Rationales for Systematic Literature Reviews (SPAR-4-SLR) framework and structured data extraction using the TCCM framework. This approach ensured comprehensive documentation of search strategies and systematic coding of the evidence base (Page et al., 2021; Paul et al., 2021; Paul & Rosado-Serrano, 2019; Rethlefsen et al., 2021).

Eligibility criteria

We included empirical studies that met the following criteria:

- Population and intervention: Involved ChatGPT or compatible LLM-based chatbots used by undergraduate English language majors (or EFL/ESL undergraduates) for writing and/or speaking development.

- Outcomes: Reported learning outcomes (e.g., accuracy, fluency, complexity, coherence for writing; pronunciation/prosody, fluency, interactional competence for speaking), perceptions linked to learning, or assessment quality.

- Publication type: Were peer-reviewed journal articles, conference papers, or accepted “advance online” articles published in English.

We excluded kindergarten through 12th grade (K-12) samples, purely conceptual or policy pieces without empirical data, and studies centered on translation only.

Information sources and search strategy

We searched five core academic databases: Scopus, Web of Science Core Collection (WoS), ERIC (EBSCOhost), EBSCO Academic Search, and SSRN. Search construction and documentation adhered to the PRISMA-S guidance (Rethlefsen et al., 2021).

- Scopus (TITLE-ABS-KEY): (chatgpt OR "large language model*" OR "generative ai") AND (EFL OR ESL OR "english language major*" OR TESOL OR "applied linguistics") AND (writing OR "academic writing" OR "L2 writing" OR speaking OR pronunciation OR "oral proficiency") AND ("higher education" OR undergraduate).

- WoS (Topic/TS): TS = (chatgpt OR "large language model*" OR "generative ai") AND TS=(EFL OR ESL OR "english language major*" OR TESOL OR "applied linguistics") AND TS=(writing OR "academic writing" OR "L2 writing" OR speaking OR pronunciation OR "oral proficiency") AND TS=("higher education" OR undergraduate)

- ERIC/EBSCO (TI, AB): (TI chatgpt OR AB chatgpt OR TI "large language model*" OR AB "large language model*" OR TI "generative ai" OR AB "generative ai") AND (TI EFL OR AB EFL OR TI ESL OR AB ESL OR TI "english language" OR AB "english language") AND (TI writing OR AB writing OR TI speaking OR AB speaking OR TI pronunciation OR AB pronunciation) AND (TI undergraduate OR AB undergraduate OR TI "higher education" OR AB "higher education")

- SSRN (All fields): chatgpt AND (EFL OR ESL OR "english language") AND (writing OR speaking).

A total of 708 records were identified (676 extracted from search queries and 32 from citation strings), from which 31 studies were ultimately included for analysis.

RESULTS

Study corpus and designs

Across the screened literature, empirical work on ChatGPT in ESL/EFL higher education has expanded rapidly, with a field-level review cataloguing 70 empirical studies within 18 months of release and noting concentration in Asia, first-year writing courses, and mixed small-N classroom designs (Lo et al., 2024). Calls from applied linguistics and second language acquisition (SLA) emphasize that this growth outpaces rigorous causal designs and theory-driven mechanism testing, urging systematic research connecting pedagogical mechanisms to outcomes (Fang & Han, 2025; Han, 2024).

The included corpus spans multiple regions and designs (Table 1). Notably, East Asia and Pacific dominates the empirical subset, which constrains global generalizability. Therefore, we interpret effects as context-bound and highlight transferability conditions rather than universal claims.

Table 1. Region and context distribution

Source	Region	Discipline	Level	Sample size	Duration	Design
Tsai et al. (2024)	Taiwan (East Asia and Pacific)	Applied Linguistics / EFL writing	Undergraduate English majors (EFL)	44 students; 44 original + 44 revised essays	Single session (paired pre-post, Feb 2023)	Prospective double-blinded paired-comparison; randomized cross-over grading
Teng (2024)	Macao SAR, China (East Asia and Pacific)	EFL writing / perceptions	Undergraduate EFL learners	≈ 90 survey + 8 interviews	Cross-sectional survey + interviews (course-based)	Mixed methods (quantitative survey; qualitative interviews)
Tram et al. (2024)	Viet Nam (East Asia and Pacific)	EFL self-directed learning / acceptance	Undergraduate EFL learners	344 survey + 19 interviews; + SLR (40 studies)	Cross-sectional survey + interviews	Multi-methods (systematic review + survey + interviews)
Tarchi et al. (2025)	Italy (Europe and Central Asia)	Educational psychology / source-based writing	First-year psychology undergraduates	27 participants (M_age≈20.4)	Timed single session (2023)	Exploratory experiment + retrospective think-aloud
Uchida (2024)	Japan (East Asia and Pacific) [dataset multi-Asia]	Assessment / writing and speaking scoring	Not applicable (secondary corpus; ICNALE GRA)	ICNALE GRA: 140 writing + 140 speaking samples; 80 raters	Not applicable (secondary analysis)	Verification study (correlation/validity)
Üstünbaş (2024)	Türkiye (Europe and Central Asia)	CALL / speaking practice	Pre-intermediate EFL undergraduates	4	8 weeks	Qualitative case study (screen recordings + interviews + stimulated recall)
Li et al. (2024)	China (East Asia and Pacific)	Assessment / reliability and feedback relevance	Undergraduate non-English majors (CET-4 essays)	30 essays; 4 teacher raters; ChatGPT-3.5/4	Single class session (30 min)	G-theory reliability + qualitative feedback analysis
Lu et al. (2024)	China (East Asia and Pacific)	Assessment / AI-assisted teacher assessment	Undergraduates (Chinese academic writing)	46 undergraduates	Course-embedded tasks (NR)	Mixed-methods (ICC consistency + interviews)
Cummings et al. (2024)	United States (North America)	Composition / first-year writing	Higher education (FYW)	Not reported (conceptual/early classroom analyses)	Not applicable	Analytical/position paper with early classroom analyses
Jiang et al. (2024)	China (dataset from a large-scale assessment)	Assessment integrity / detector bias	Large-scale writing assessment (operational)	NR (large-scale operational dataset)	Not applicable	Observational fairness analysis (detector bias vs NNS)
Jiang et al. (2024)	China (East Asia and Pacific)	Assessment integrity / keystroke and fairness	Writing assessment (lab/operational)	NR	Not applicable	Predictive modelling of non-authentic texts; subgroup fairness

Liang et al. (2023)	United States (North America)	AI detection / bias toward NNS	Not applicable	Benchmark datasets (NR)	Not applicable	Algorithmic audit / bias analysis
Zare et al. (2025)	Iran/UAE/Oman (Middle East)	EFL argumentative writing / motivation	Undergraduates (L2 learners)	≈ 69 (experimental + control; mixed-methods)	Pre-test → immediate post-test → 1-month delayed post-test	Randomized pre-/post-test control group + mixed-methods

Writing outcomes: performance and assessment

Across undergraduate contexts, writing performance improves when ChatGPT is integrated into drafting and revision cycles, particularly under human-in-the-loop orchestration. A double-blinded paired-comparison with EFL English majors found that ChatGPT-assisted revisions significantly increased scores on four dimensions (largest gains in vocabulary), while also raising fairness concerns because benefits were disproportionately large for initially weaker writers (Tsai et al., 2024). Reliability studies converge: ChatGPT shows moderate-to-good consistency with teacher scoring (intraclass correlations), expands feedback volume/types (Lu et al., 2024), and with G-theory ChatGPT-4 achieved higher reliability coefficients than teachers in CET-4 essay scoring and provided more relevant qualitative feedback than teachers on language/content/organization (Li et al., 2024).

Learner-centered work complements these findings. Students who position ChatGPT as a "companion" reported increased accuracy and coherence during drafting/revision, alongside positive perceptions of iterative feedback (Teng, 2024). For task motivation, experimental evidence indicates that interacting with ChatGPT can raise motivation to write argumentative essays, though effects may attenuate at delay, signaling novelty and self-regulation issues (Zare et al., 2024). At the same time, in source-based writing, undergraduates often use ChatGPT non-strategically and show weaker integration of literal source content; ethical concerns and low prompting expertise suppress usage (Tarchi et al., 2024). These patterns reaffirm that prompting scaffolds, rubrics, and teacher moderation are central for higher-order writing development beyond surface correctness (Cummings et al., 2024; Lo et al., 2024).

A broader meta-analytic lens suggests that carefully designed ChatGPT interventions improve academic performance ($g^+ \approx 0.71$), higher-order thinking ($g^+ \approx 0.70$), and affective-motivational states ($g^+ \approx 0.88$), while reducing mental effort (Deng et al., 2025). Heterogeneity remains large, implying sensitivity to subject area, duration, and application mode, and highlighting a need for moderator-aware designs in L2 writing (Deng et al., 2025).

Speaking outcomes: practice, affect, and measurement gaps

Evidence for speaking is emergent but promising. Case-based and small-N studies report that students use ChatGPT to rehearse dialogues, receive corrective feedback, and lower anxiety in low-stakes practice (Üstünbaş, 2024). Reviews and state-of-the-field papers nonetheless stress that oral proficiency measurement lags behind writing, calling for CEFR-aligned rubrics, ASR-derived features (e.g., speech rate, pause ratio), and stronger validation (Fang & Han, 2025; Lo et al., 2024). Initial verification work using the International Corpus Network of Asian Learners of English - Global Rating Archive (ICNALE GRA) corpus finds high correlations between ChatGPT and human ratings for writing and exploratory alignment for speaking assessment, useful but not a substitute for validated oral scoring protocols (Uchida, 2024). Overall, practice opportunities scale readily, but pronunciation/prosody and interactional competence remain under-measured in current trials.

Table 2 synthesizes speaking-assessment practices across the included studies, detailing the speaking task type, the extent of CEFR alignment, ASR-derived acoustic indicators (e.g., speech rate, articulation rate, pause ratio, mean length of run, pitch variability), inter-rater reliability (ICC/κ), intervention duration, sample size, and study design. The evidence indicates substantial fragmentation: CEFR-based rubrics are not widely adopted, ASR metrics and reliability estimates are reported only in a minority of studies, and designs are frequently short and under-powered - limitations that reduce cross-study comparability and weaken inference. We therefore propose a minimum reporting toolkit for future research: (i) standardized CEFR-aligned rubrics; (ii) a core set of ASR indices capturing fluency and prosody; and (iii) mandatory reporting of ICC/κ to enhance reproducibility and enable more rigorous evidence synthesis.

Table 2. Speaking measurement practices in the included corpus

Study	Region	Level	Speaking task	Rubric	Inter-rater reliability	Design	Duration	Notes
Üstünbaş (2024)	Türkiye (Europe and Central Asia)	Pre-intermediate EFL undergraduates	ChatGPT-mediated speaking practice (guided conversation)	Not reported	Not reported	Qualitative case study	8 weeks	Screen recordings + stimulated recall; perceived gains; limited measurement rigor
Uchida (2024)	Japan (dataset multi-Asia)	N/A (secondary dataset; speaking and writing samples)	Standardized short responses (ICNALE)	ICNALE scoring; CEFR mapping possible (not explicit)	80 human raters; validity/ consistency checks	Verification study (correlation between human and model ratings)	N/A	Writing–speaking rating comparison; speaking correlations lower than writing

Table 3. Integrity approaches & fairness considerations

Approach	Representative studies	Evidence summary	Key risks (fairness)	Data/infra requirements	Classroom feasibility	Implementation notes
Text-only AI detectors	Liang et al. (2023); Jiang et al. (2024)	Detectors exhibit brittleness and false positives, especially for non-native writers; high variance across prompts/models.	Penalizing NNS; domain/topic bias; version drift; opacity	Detector access; reference corpora; threshold tuning	Low–moderate (licensing and false positives limit adoption)	Use only as triage; never sole evidence for misconduct
Keystroke/ behavioral analytics	Jiang et al. (2024)	Temporal/pausal dynamics can flag non-authentic authorship with better specificity when validated.	Privacy; accessibility accommodations; device heterogeneity	Logging tools; consent; secure storage; calibration by cohort	Moderate (needs LMS/editor integration)	Combine with reflective artefacts; clear consent & ethics
Process-anchored integrity (recommended)	Cummings et al. (2024); Lo et al. (2024)	In-class drafting + version histories + prompt logs + short viva provides transparent authorship evidence while maintaining learning value.	Workload; training for staff; consistency across sections	Versioned docs; prompt logging; viva rubric	High (policy-aligned; pedagogy-friendly)	Adopt 'human-in-the-loop'; publish rubric and due-process steps

Integrity, fairness, and process-based assessment

The integrity landscape is complex. Large-scale evidence in computers and education suggests detector behavior can misclassify texts and raises questions of bias against non-native English speakers in operational settings (Jiang et al., 2024). Foundational work in Patterns shows that popular detectors systematically flag non-native speakers (NNS) writing as AI-generated while showing near-perfect accuracy for native samples, underscoring risks in evaluative settings (Liang et al., 2023). A complementary strand uses keystroke dynamics to detect nonauthentic texts with attention to fairness across subgroups, suggesting a shift from product-only to process-based integrity (Jiang et al., 2024). Taken together, institutions are moving towards in-class drafting, oral verification, and prompt logs/version histories rather than relying on brittle text detectors (Cummings et al., 2024; Lo et al., 2024).

Table 3 contrasts three families of integrity safeguards for generative-AI use in undergraduate EFL writing/ speaking: (i) text-only AI detectors; (ii) keystroke/behavioral analytics; and (iii) process-anchored integrity (in-class drafting, version histories/prompt logs, brief viva). For each approach, the table synthesizes the evidence base, typical failure modes and fairness risks (e.g., false positives for non-native writers), data/infra prerequisites, classroom feasibility, and implementation notes. The comparison indicates that detector-only solutions remain brittle and potentially inequitable, behavioral analytics can improve specificity but raise privacy/accessibility concerns, whereas process-anchored designs provide the strongest, pedagogically aligned evidence of authorship. Table 3 motivates our stance: use detectors only as triage signals, employ behavioral traces with explicit consent and safeguards, and adopt process-anchored workflows as the default integrity regime.

Figure 1 illustrates a process-anchored integrity workflow for undergraduate EFL writing/speaking creating a chain of evidence that authenticates authorship while enhancing equity and learning value.

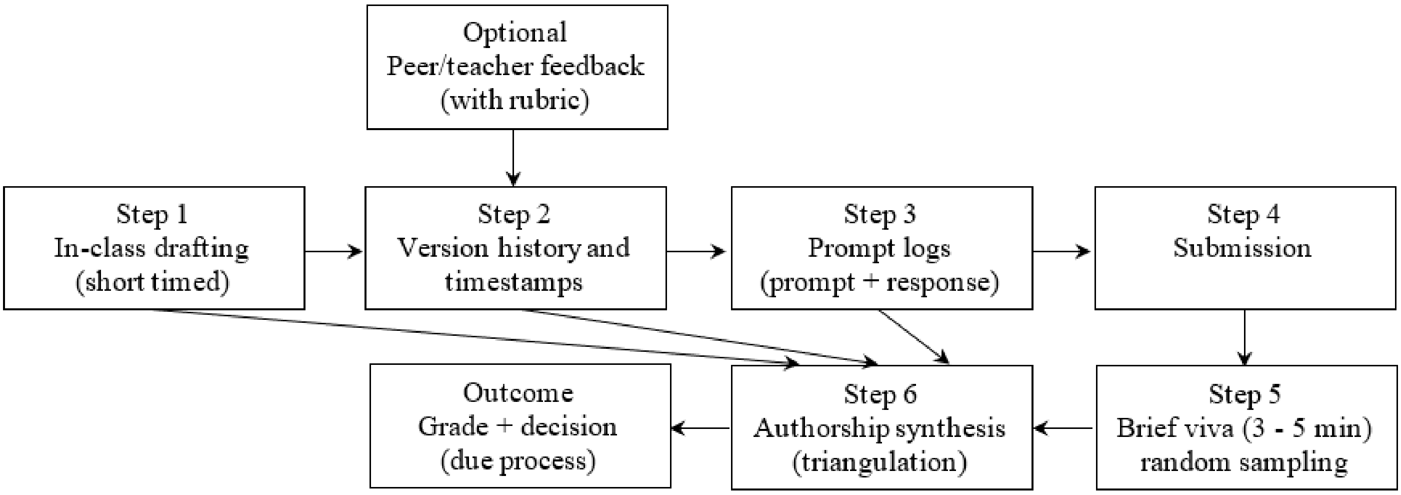

Figure 1. Process-anchored integrity workflow for undergraduate EFL writing/speaking

Moderators and patterns

Across the included studies, student adoption of ChatGPT was generally positive, with perceived gains in efficiency and reduced speaking anxiety. Acceptance tended to be higher when instructor-provided prompts and checkpoints were present, and lower in large-cohort contexts without clear guidance. Instructors reported benefits for formative feedback and drafting support but raised concerns about over-reliance and authorship evidence.

Following SWiM guidance for narrative syntheses, we group studies by pre-specified moderators and report directionality of patterns (↑ improvement; ↔ mixed/no clear change; ↓ deterioration) rather than pooled effects. The patterns below are indicative given design heterogeneity and short intervention durations and should guide future moderator-aware RCTs and meta-analyses. These descriptive patterns are summarized here to characterize the learning contexts; their moderating roles are synthesized in Table 4.

Table 4. Moderators and observed patterns across the included studies

Moderator	**Operationalization**	**Observed pattern**	**Illustrative evidence**	**Implications for practice**
Task type	Argumentative/expository; source-based synthesis; narrative/personal response; speaking rehearsal/role-play	Writing: ↑ for argumentative/ source-based under scaffolded prompting; Narrative/free-response ↔. Speaking: affect (anxiety) ↑; proficiency ↔ without rigorous measurement.	Tarchi et al. (2025) source-based writing; Tsai et al. (2024) argumentative writing; Üstünbaş (2024) speaking practice.	Prioritize structured tasks with explicit criteria; design speaking tasks with measurable sub-skills and standardized rubrics.
Modality and toolchain	Text-only chat; multimodal planning + drafting; speech practice with ASR feedback	Text writing ↑ (quality/accuracy); Multimodal planning ↑ if human moderation present; Speaking proficiency ↔ (measurement limits), affect ↑.	Li et al. (2024) writing reliability; Lo et al. (2024) ESL/EFL overview; Üstünbaş (2024) speaking.	Integrate ASR metrics for speaking; couple chat with versioned documents for traceability.
Human oversight (human-in-the-loop)	Unsupervised use; prompt scaffolds; iterative feedback with teacher moderation	With scaffolds/moderation ↑ (writing quality, process transparency); unsupervised ↔/↓ (surface-level paraphrase, over-reliance).	Cummings et al. (2024) classroom orchestration; Lo et al. (2024) practice guidance.	Make teacher prompts and checkpoints explicit; require process artefacts to anchor authorship.
Baseline proficiency	A2 - B1 (lower-intermediate); B1 - B2 (intermediate); C1 (advanced)	B1–B2 learners: ↑ (largest practical gains). A2: ↔ without heavy scaffolding. C1: ↔ (diminishing marginal returns).	Teng (2024) perceptions; Tram et al. (2024) multi-methods; Tsai et al. (2024).	Tailor scaffolds by proficiency; pre-register moderator analyses by CEFR bands.
Feedback/ prompting literacy	No training; brief prompt hygiene; structured prompting + feedback literacy curriculum	Structured prompting + feedback literacy ↑; minimal training ↔.	Cummings et al. (2024); Lo et al. (2024).	Teach prompt patterns and feedback uptake; capture logs for formative analytics.
Dose and duration	Single session; 4 - 8 weeks; ≥ 12 - 16 weeks (multi-site)	Single session ↔ (transient gains). 4 - 8 weeks ↑ modest. ≥ 12 - 16 weeks expected ↑ and more durable (rare in current corpus).	Deng et al. (2024) meta-analytic lens; Üstünbaş (2024) 8-week case.	Plan multi-week designs; pre-specify retention/ delayed post-tests.
Assessment regime	Detector-led (text-only); keystroke/behavioral traces; process-anchored (versions + logs + viva)	Process-anchored ↑ (learning + integrity). Detector-only ↔/↓ (brittleness, equity risks).	Cummings et al. (2024); Jiang et al. (2024); Liang et al. (2023).	Adopt process-anchored integrity by default; use detectors only as triage with due process.
Class size/ instructor bandwidth (contextual)	Small seminar; medium lecture-lab; large lecture	Smaller classes → ↑ with closer moderation; large cohorts ↔ unless workflows are streamlined.	Programmatic observations across classroom studies.	Standardize workflows (checkpoints, rubrics, templates) to scale human-in-the-loop.

Taken together, Table 4 suggests that the largest and most defensible gains arise when (i) tasks are structured and criteria-rich, (ii) human-in-the-loop scaffolds are explicit, (iii) learners are at B1 - B2 proficiency with prompt/feedback literacy training, and (iv) integrity is anchored in documented process rather than detector scores. Conversely, single-session, unsupervised uses in open-ended tasks produce at best transient improvements and weaker authorship evidence. These patterns imply concrete design choices for future trials: moderator-stratified sampling, ≥ 12 - 16-week interventions, delayed post-tests for retention, and pre-registered moderator analyses (e.g., task type × human oversight × proficiency).

Moderators and boundary conditions

Patterns across studies indicate several moderators:

- Task type: benefits are clearest for iterative revision and argumentative writing; source-based synthesis demands explicit prompting and source-use scaffolds (Tarchi et al., 2024).
- Modality and orchestration: text-only chat supports writing mechanics; voice modes scale rehearsal but require latency-aware designs and validated oral rubrics (Fang & Han, 2025; Üstünbaş, 2024).
- Human oversight: teacher/peer mediation consistently strengthens learning and mitigates over-reliance (Li et al., 2024; Lu et al., 2024; Teng, 2024).
- Learner factors: initial proficiency and feedback literacy condition transfer from surface accuracy to coherence/argument quality (Lo et al., 2024; Zare et al., 2024).

In aggregate, undergraduate EFL writing benefits most when ChatGPT is embedded in scaffolded, monitored processes that emphasize revision and rhetorical development; speaking benefits are visible in practice and affect, but robust measurement remains a gap. Integrity and fairness concerns caution against detector-led policing and favor process-based assessment. Heterogeneity in effects calls for moderator-aware designs and validated outcome measures moving forward (Deng et al., 2025; Han, 2024; Lo et al., 2024).

DISCUSSION AND LIMITATIONS

This discussion integrates the study corpus to explain when and why ChatGPT assists undergraduate English majors, and what infrastructural and pedagogical conditions are required for robust, equitable gains. Across the evidence, writing improvements are most defensible when human moderation and structured task design are present, whereas speaking benefits are presently concentrated in affect (e.g., reduced anxiety) with limited proficiency gains due to measurement gaps. Integrity safeguards based solely on text-only detectors remain brittle and raise fairness concerns for non-native writers; a process-anchored regime is preferable. Generalizability is constrained by an Asia-leaning empirical base and the prevalence of short, small-sample studies. We therefore frame recommendations on two axes: (i) technical infrastructure and measurement, and (ii) pedagogical orchestration.

Technical infrastructure and measurement

Standardizing outcome measurement is the most immediate lever for cumulative knowledge. For writing, recent studies indicate that model-assisted ratings can approach or exceed the reliability of human raters when anchored to explicit rubrics and used with human moderation (e.g., Li et al., 2024; Lu et al., 2024). For speaking, however, few studies report CEFR-aligned rubrics, acoustic indicators derived from automatic speech recognition (ASR) or inter-rater reliability (ICC/κ). We propose a minimum reporting toolkit: CEFR-aligned rubrics (B1 - C1), a core ASR bundle for fluency/prosody (speech rate, articulation rate, pause ratio, mean length of run, pitch variability), and mandatory reporting of inter-rater reliability. Table 2 summarizes current practice gaps and motivates this toolkit.

Second, integrity evidence should be grounded in production processes rather than post-hoc text classification. Detector-only approaches are prone to false positives and subgroup disparities for non-native writers (Liang et al., 2023; Jiang et al., 2024). By contrast, process-anchored integrity triangulates in-class drafting, version histories with timestamps, prompt/response logs, and a brief viva for a sampled subset; this yields clearer authorship evidence while preserving learning value (Cummings et al., 2024; Lo et al., 2024). Table 3 and Figure 1 operationalize this workflow.

Third, study design and reporting should reflect contemporary evidence standards. Where meta-analysis is infeasible, narrative synthesis should follow SWiM guidance, with pre-specified groupings (e.g., task type × human oversight × proficiency) and clear rules for directionality coding. Risk-of-bias appraisals (RoB 2; ROBINS-I) should be reported, and screening/reporting should adhere to PRISMA-2020/PRISMA-S. Power analysis, multi-site trials of ≥ 12 - 16 weeks, and delayed post-tests are necessary to move beyond transient, under-powered findings (see Deng et al., 2024/2025 for a cross-domain meta-analytic lens).

Finally, data governance and ethics: behavioral traces (e.g., keystrokes) require informed consent, clear retention policies, and accessibility accommodations; detectors - if used at all - serve only as triage signals and never as sole evidence. These safeguards should be codified in course-level integrity statements and rubric-linked due-process procedures.

Pedagogical orchestration

The largest and most defensible gains arise when ChatGPT is orchestrated as part of a transparent, scaffolded writing-and-speaking workflow. For writing, structured prompts and human-in-the-loop moderation shift usage from surface paraphrase to higher-order planning, argumentation, and revision. Evidence from classroom implementations suggests improvements in accuracy, coherence, and organization under scaffolded prompting, with teacher checkpoints providing both formative feedback and integrity anchors (Cummings et al., 2024; Tsai et al., 2024). Source-based tasks benefit particularly from explicit criteria and staged drafting (Tarchi et al., 2025).

For speaking, current benefits are strongest for affective outcomes (confidence, reduced anxiety) and for rehearsal/role-play contexts, but proficiency effects remain under-measured. Pedagogical design should therefore combine low-stakes practice with measurable sub-skills (interaction, fluency, pronunciation), integrate ASR-supported feedback where feasible, and adopt CEFR-aligned rubrics with rater calibration. Where class sizes are large, standardized checkpoints (brief in-class drafting, versioned submissions, short viva sampling) help scale human oversight without resorting to punitive detectors (Uchida, 2024; Üstünbaş, 2024).

Learner characteristics and enabling conditions matter. B1 - B2 learners appear to realize the largest practical gains; A2 learners require heavier scaffolding; advanced learners exhibit diminishing returns without higher-order task demands (Teng, 2024; Tram et al., 2024). A brief curriculum in prompting and feedback literacy improves uptake and mitigates over-reliance. At the programmed level, departments should standardize artifacts (prompt templates, version-control norms, integrity rubrics) to ensure parity of experience across sections and to support moderator-aware evaluation (see Table 4).

In short, the pedagogical stance is not “AI or human,” but human-in-the-loop with transparent processes. When aligned with rigorous measurement and integrity infrastructure, this orchestration can deliver durable writing gains and a more equitable environment for speaking practice, while enabling cumulative research that is transferable beyond the current Asia-leaning evidence base (Cummings et al., 2024; Lo et al., 2024).

Limitations

In this review, we used a comprehensive strategy consistent with PRISMA-2020 and PRISMA-S to document sources and search strings; however, study-level heterogeneity precluded a pooled effect size, so we adopted a structured narrative synthesis. Screening and extraction procedures were standardized (e.g., Rayyan) and quality/risk-of-bias tools were consulted (ROBINS-I; RoB 2), yet residual selection and reporting biases are possible -particularly where grey literature and non-English outlets may not be fully captured. As in prior reviews of AI-assisted language learning, construct alignment across studies (task type, proficiency, orchestration, assessment criteria) remained uneven.

For the current evidence base, there are some limits:

- First, the geographical concentration of empirical studies in East Asia and Pacific and allied higher-education settings constrains external validity; transfer to other regions, program structures, and assessment regimes should be made cautiously (e.g., China, Japan, Viet Nam).

- Second, many interventions are short-duration (single-session to < 4 weeks) and/or small-sample designs (case studies, pilot RCTs), which limits statistical power and precision and increases sensitivity to context and task idiosyncrasies.

- Third, speaking proficiency is under-measured: relatively few studies report CEFR-aligned rubrics, ASR-based acoustic indicators (e.g., speech rate, articulation rate, pause ratio, mean length of run), or inter-rater reliability (ICC/κ). Evidence that does triangulate human and model ratings suggests lower correspondence for speaking than for writing.

- Fourth, the integrity/fairness corpus is still emergent. Text-only AI detectors show brittleness and false-positive risk for non-native writers; keystroke/behavioral analytics improve specificity but raise privacy and implementation concerns.

- Finally, the literature remains heterogeneous in task design, “dose” of AI use, and orchestration (human-in-the-loop vs. fully autonomous), making meta-analytic synthesis challenging; narrative synthesis following SWiM guidance mitigates but does not remove subjectivity.

CONCLUSION & RECOMMENDATIONS

Across the corpus, the most reliable gains occur in L2 writing when ChatGPT is embedded in iterative drafting/revision with teacher or peer mediation; effects are weaker or inconsistent for un-scaffolded, single-shot tasks. Benefits for speaking are visible in increased practice and reduced anxiety, yet the evidence base is constrained by fragile measurement and limited validation of CEFR-aligned oral rubrics and instrumentation. System-level concerns about integrity and equity - especially detector misclassification of non-native writing - argue for process-anchored assessment rather than detector-led policing. Meta-analytic results across disciplines are encouraging but heterogeneous, underscoring the need for moderator-aware designs. Overall, the field is promising but under-theorized and under-measured in speaking, with over-representation of short, single-site studies.

Pedagogical recommendations

Scaffolded prompting & feedback literacy. Treat prompting and feedback use as learnable skills; require brief rationales for accepting or rejecting AI suggestions, anchored to analytic rubrics.

Human-in-the-loop orchestration. Combine ChatGPT with peer/teacher conferencing; allow convergence of human and AI judgements while keeping teachers as final arbiters to maintain rhetorical development and authorial voice.

Process-based assessment for integrity. Replace detector-only policing with in-class drafting, version histories, prompt logs, and brief oral verification; these approaches both protect equity and preserve learning value.

Research recommendations.

Speaking measurement. Co-develop validated CEFR-aligned oral rubrics and ASR-derived features (speech rate, pause ratio, prosody) and report rater reliability alongside outcomes.

Causal strength and sustainability. Conduct multi-site RCTs (≥ 12 - 16 weeks), report follow-ups, and publish process data to connect mechanisms to outcomes.

Transparency and equity. Pre-register protocols, share materials/data, and audit fairness for non-native writers whenever detection or automation is used.

Table 5 summarizes actionable recommendations by stakeholder, the evidentiary basis, KPIs and timelines for implementation.

Table 5. Practical recommendations by stakeholder

Stakeholder	**Practical recommendations**	**Key risks and mitigation**	**KPIs / success indicators**	**Timeframe**	**Evidence**
Instructors (course level)	• Adopt human-in-the-loop use of ChatGPT; implement process-anchored integrity (in-class drafting, version/prompt logs, short viva) • Use scaffolded prompting templates and require ‘prompt portfolios’ (students submit prompts + model outputs + reflections) • Writing: grade with CEFR-aligned rubrics; double-mark 10 - 20% scripts to report inter-rater reliability (target ICC ≥ .75) • Speaking: add ASR indicators (e.g., speech rate, articulation rate, pause ratio, mean length of run) alongside CEFR descriptors • Design AI-resilient tasks (local data, personal evidence, oral defense)	Workload, privacy, detector false positives → Use templates/checklists; obtain consent; detectors not sole evidence; viva as safety net	• % assignments with process-anchored integrity • ICC value reported per course/assessment • % courses using CEFR-aligned rubrics (writing/speaking) • # speaking tasks with ASR indicators; change in anxiety/self-efficacy	Now (pilot this term) → Next term (scale)	Lo et al. (2024); Li et al. (2024); Lu et al. (2024); Üstünbaş (2024); Uchida (2024); Tsai et al. (2024); Tarchi et al. (2025); Liang et al. (2023); Jiang et al. (2024)
Assessment and QA units (department/ faculty)	• Publish policy: detectors not used as sole evidence; due-process anchored in process artefacts • Standardize measurement: CEFR rubrics (writing/speaking), ASR indicator set, mandatory reliability (ICC/κ) reporting • Run fairness audits for subgroup disparities and document false-positive rates	Policy acceptance; staff capacity → Provide exemplars, training, phased roll-out	• Policy published and communicated • % assessments with CEFR+ASR+ICC reporting • Fairness audit completed per semester; FP rate tracked	Now → Next term	Liang et al. (2023); Jiang et al. (2024 JEM; 2024 C&E); Lo et al. (2024)
Program administrators (curriculum leaders)	• Embed AI-literacy and feedback-literacy outcomes across writing/speaking modules • Provide micro-PD for instructors (prompting, assessment with CEFR/ASR, fairness) • Ensure cross-section consistency (common rubrics, shared task bank, moderation)	Fragmentation across sections → Common assessment frameworks; moderation panels	• % modules with explicit AI/feedback-literacy LOs • PD coverage (% instructors trained) • Moderation records; variance of grades across sections	Next term → Next academic year	Lo et al. (2024); Deng et al. (2025) (cross-domain gains under structured orchestration)
Students (undergraduate	• Complete orientation on ethical use, prompt design, and reflective practice	Over-reliance on AI → Reflection prompts; oral	• Orientation completion rate; quality of process	Now	Üstünbaş (2024); Lo et al. (2024); Uchida

English majors)	• Maintain personal 'process dossiers' (drafts, version histories, prompt logs, reflections) • Use ChatGPT to plan/practice speaking with self-monitoring (ASR-based dashboards when available)	defense; teacher feedback cycles	dossiers • Hours of speaking practice; improvement in fluency/prosody proxies • Self-efficacy/anxiety change (pre–post)		(2024)
EdTech / IT (infrastructure and privacy)	• Provide versioned editors and prompt-logging integrated with LMS (opt-in; consent-based) • Enable ASR plug-ins for speaking metrics; ensure secure storage and minimal retention • Implement consent management and access controls; run privacy impact assessments	Privacy, security, interoperability → PIA, DPA compliance, vendor due diligence, pilot before scale	• Tools deployed and uptime • % courses using process logging; privacy incidents = 0 • Student consent records; ASR data coverage	Next term → Next academic year	Jiang et al. (2024); Lo et al. (2024)
Researchers (faculty/grad labs)	• Run multi-site RCTs (≥ 12 - 16 weeks) with pre-registration; adopt PRISMA-2020/SWiM in syntheses • Publish measurement toolkits: CEFR-aligned rubrics; minimal ASR indicator set; reliability reporting templates • Share data/code (OSF/Zenodo); audit fairness across subgroups	Resource intensity → Consortia, shared protocols, seed grants	• # pre-registered RCTs; study duration • Open materials/data availability • Completeness of measurement (CEFR+ASR+ICC)	Next academic year	Page et al. (2021); Campbell et al. (2020); Deng et al. (2025)

ACKNOWLEDGMENT

This study was conducted as part of the student scientific research project code THS 2025-15. The authors would like to express their sincere gratitude to Can Tho University for funding and supporting this study.

REFERENCES

1. Aromataris, E., Lockwood, C., Porritt, K., Pilla, B., & Jordan, Z. (Eds.). (2024). JBI manual for evidence synthesis. JBI. https://jbi-global-wiki.refined.site/space/MANUAL
2. Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Introduction to meta-analysis. Wiley.
3. Campbell, M., McKenzie, J. E., Sowden, A., Katikireddi, S. V., Brennan, S. E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, J., & Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. BMJ, 368, l6890. https://doi.org/10.1136/bmj.l6890
4. Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46. https://doi.org/10.1177/001316446002000104
5. Cummings, R. E., Monroe, S. M., Watkins, M. (2024). Generative AI in first-year writing: An early analysis of affordances, limitations, and a framework for the future. Computers and Composition, 71, 102827. https://doi.org/10.1016/j.compcom.2024.102827
6. Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. Computers & Education, 227, 105224. https://doi.org/10.1016/j.compedu.2024.105224
7. Fang, S., & Han, Z. (2025). On the nascency of ChatGPT in foreign language teaching and learning. Annual Review of Applied Linguistics, 45, 148–178. https://doi.org/10.1017/S026719052510010X
8. Han, Z. (2024). ChatGPT in and for second language acquisition: A call for systematic research. Studies in Second Language Acquisition, 46(2), 301–306. https://doi.org/10.1017/S0272263124000111

9. Jiang, Y., Hao, J., Fauss, M., & Li, C. (2024). Detecting ChatGPT-generated essays in a large-scale writing assessment: Is there a bias against non-native English speakers? Computers & Education, 217, 105070. https://doi.org/10.1016/j.compedu.2024.105070
10. Jiang, Y., Zhang, M., Hao, J., Deane, P., & Li, C. (2024). Using keystroke behavior patterns to detect nonauthentic texts in writing assessments: Evaluating the fairness of predictive models. Journal of Educational Measurement, 61(4), 571–594. https://doi.org/10.1111/jedm.12431
11. Li, J., Huang, J., Wu, W., & Whipple, P. B. (2024). Evaluating the role of ChatGPT in enhancing EFL writing assessments in classroom settings: A preliminary investigation. Humanities & Social Sciences Communications, 11, 1268. https://doi.org/10.1057/s41599-024-03755-2
12. Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. Patterns, 4(7), 100779. https://doi.org/10.1016/j.patter.2023.100779
13. Lo, C. K., Yu, P. L. H., Xu, S., Ng, D. T. K., & Jong, M. S. Y. (2024). Exploring the application of ChatGPT in ESL/EFL education and related research issues: A systematic review of empirical studies. Smart Learning Environments, 11, 50. https://doi.org/10.1186/s40561-024-00342-5
14. Lu, Q., Yao, Y., Xiao, L., & Yuan, M. (2024). Can ChatGPT effectively complement teacher assessment of undergraduate students' academic writing? Assessment & Evaluation in Higher Education. Advance online publication. https://doi.org/10.1080/02602938.2024.2301722
15. McHugh, M. L. (2012). Interrater reliability: The kappa statistic. Biochemia Medica, 22(3), 276–282. https://www.biochemia-medica.com/en/journal/22/3/10.11613/BM.2012.031/fullArticle
16. Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—A web and mobile app for systematic reviews. Systematic Reviews, 5, 210. https://doi.org/10.1186/s13643-016-0384-4
17. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71
18. Paul, J., Lim, W. M., & O'Cass, A. (2021). Scientific procedures and rationales for systematic literature reviews (SPAR-4-SLR). International Journal of Consumer Studies, 45(6), 1513–1527. https://doi.org/10.1111/ijcs.12695
19. Paul, J., & Rosado-Serrano, A. (2019). Gradual internationalization vs born-global/ international new venture models: A review and research agenda. International Marketing Review, 36(6), 830–858. https://doi.org/10.1108/IMR-10-2018-0280
20. Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., & Koffel, J. B. (2021). PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. Systematic Reviews, 10, 39. https://doi.org/10.1186/s13643-020-01542-z
21. SCImago (n.d.). SJR—SCImago Journal & Country Rank [Portal]. Retrieved October 12, 2025, from https://www.scimagojr.com
22. Sterne, J. A. C., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M.,… & Higgins, J. P. T. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. BMJ, 355, i4919. https://doi.org/10.1136/bmj.i4919
23. Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., … & Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. BMJ, 366, l4898. https://doi.org/10.1136/bmj.l4898
24. Tarchi, C., Zappoli, A., Casado Ledesma, L., & Wennås Brante, E. (2025). The use of ChatGPT in source-based writing tasks. International Journal of Artificial Intelligence in Education, 35, 858–878. https://doi.org/10.1007/s40593-024-00413-1
25. Teng, M. F. (2024). "ChatGPT is the companion, not enemies": EFL learners' perceptions and experiences in using ChatGPT for feedback in writing. Computers & Education: Artificial Intelligence, 7, 100270. https://doi.org/10.1016/j.caeai.2024.100270
26. Tram, N. H. M., Nguyen, T. T., & Tran, C. D. (2024). ChatGPT as a tool for self-learning English among EFL learners: A multi-methods study. System, 127, 103528. https://doi.org/10.1016/j.system.2024.103528
27. Tsai, C. Y., Lin, Y. T., & Brown, I. K. (2024). Impacts of ChatGPT-assisted writing for EFL English majors: Feasibility and challenges. Education and Information Technologies, 29(17), 22427–22445. https://doi.org/10.1007/s10639-024-12722-y
28. Uchida, S. (2024). Evaluating the accuracy of ChatGPT in assessing writing and speaking: A verification

study using ICNALE GRA. Learner Corpus Studies in Asia and the World, 6, 1–12. https://doi.org/10.24546/0100487710

29. Üstünbaş, Ü. (2024). Hey, GPT, can we have a chat? A case study on EFL learners' AI speaking practice. International Journal of Modern Education Studies, 8(1), 91–107. https://doi.org/10.51383/ijonmes.2024.318
30. Yang, L., & Li, R. (2024). ChatGPT for L2 learning: Current status and implications. System, 124, 103351. https://doi.org/10.1016/j.system.2024.103351
31. Zare, J., Al-Issa, A., & Ranjbaran Madiseh, F. (2025). Interacting with ChatGPT in essay writing: A study of L2 learners' task motivation. ReCALL, 37(3), 385–402. https://doi.org/10.1017/S0958344025000035