Journal of Teaching Innovation and Reform

jtirjournal.com

Research Article

Artificial Intelligence in Teaching Chinese as A Target Second Foreign Language: A Scoping Review (2021–2026)

Huixin Zong[1] and Shuqing Gao[2,*]

[1]School Of International Communications, Jilin International Studies University; E-mail: 18829810080@163.com

[2]School Of International Communications, Jilin International Studies University; E-mail: 28901260@qq.com

(*Corresponding Author)

Cite This: https://doi.org/10.65638/2978-5634.2026.2.05

ABSTRACT: Artificial intelligence is reshaping language pedagogy, but Chinese as a target, second, or foreign language remains under-synthesised in high-indexed empirical literature. This PRISMA-ScR scoping review maps 33 Chinese-target AI-related empirical studies published in JCR 2023–2024 SSCI/SCIE Q1 or Q2 venues between 2021 and May 2026. A database search supplemented by four citation-chasing records identified 248 records, 206 after deduplication, and 61 records for full-text and venue verification. The final primary set shows a three-pillar evidence problem rather than simple scarcity. First, writing AI feedback (7/33, 21.2%) and teacher-facing or practitioner studies (9/33, 27.3%) are the largest clusters, while reading has one mediation-correlational study, Chinese character / handwriting acquisition has no Q1/Q2 entry, and no direct K-12 learner intervention survives venue verification. Second, design maturity is uneven: only one true RCT is identified, 19 of 33 studies use samples below 100, and the set contains no replication, pre-registered protocol, public de-identified dataset, or L2-Chinese latent-growth-curve analysis. Third, ethics, originality, and cognitive dependence remain under-consolidated: five studies treat these as primary outcomes, but they use different constructs and instruments. The review translates these gaps into a prioritised R1–R7 agenda for Chinese-target AI-supported pedagogical innovation.

Keywords: Artificial intelligence, Generative AI, Chinese as a foreign language, Scoping review, Pedagogical innovation, PRISMA-ScR.

INTRODUCTION

The Rise of AI in Language Pedagogy

The 2021–2026 period has been an inflection in the use of AI in language pedagogy. Two technological trajectories converged. First, large language models matured into instructionally usable systems: the public release of ChatGPT in late 2022, followed by GPT-4 in 2023, GPT-4o and competitive systems in 2024–2025, made fluent multi-turn interaction in Chinese broadly accessible to teachers and learners outside specialised laboratories. Second, the surrounding stack matured in parallel: automatic speech recognition coupled with learner-model loops, voice-mode generative AI, multimodal generation (DALL-E, Sora), and educational robotics with AI components all became deployable in classroom and self-study settings.

The corresponding empirical-research output has tracked this trajectory. A systematic review and meta-analysis of AI-enabled assessment in language learning across 2012–2024 (Chen *et al*. 2025) and a scoping review of generative AI in language education across 2022–2024 (Anonymous 2025) both document that most of the empirical literature in this period addresses generative AI as a mediator of language learning, that writing is the most frequently studied pedagogical task, and that English is by a large margin the most frequently studied target language. A complementary three-level meta-

Received: May 07, 2026
Accepted: June 08, 2026
Published: June 13, 2026

analysis of dialogue-based computer-assisted language learning systems for L2 speaking, covering 16 studies and 89 effect sizes, estimates an overall pooled effect of Hedges' g = .61, 95% CI [.34, .89], and identifies system type, meaning constraint, and system modality – not learner proficiency – as the dominant moderators (Hou and Min 2026). Notably, the meta-analysis's included-study table reports no identifiable Chinese-target primary study among the 16 included investigations – a direct empirical signal that the field-level speaking-development evidence has not yet integrated Chinese-as-target-language work, which we return to in Section 3.10.

This trajectory motivates the reform-relevant question that drives the present review. As the technology and the surrounding evidence base have matured, the pedagogical question has shifted from whether AI can support language teaching to which configurations of AI, task, and learner are supported by Q1/Q2 evidence, and which remain pedagogically unexamined. For Chinese as a target language, that question is the central concern of this review.

Why Chinese as a Target Language

Pedagogical adaptations are language-specific, and the AI affordances that matter for one target language are not necessarily those that matter for another. Chinese as a target language exposes four such demands. First, Mandarin is tonal: lexical contrast and discourse prosody are carried by four citation tones plus a neutral tone, so AI-mediated pedagogy for tone turns on ASR + learner-model loops with contour-level visual or acoustic feedback – a configuration with no equivalent in non-tonal targets. Second, the writing system is logographic, requiring acquisition of several thousand characters by stroke, radical, and structural position; the relevant AI affordances (handwriting recognition with feedback, generative stroke decomposition, character-model-aware spaced repetition) are distinct from those for alphabetic literacies. Third, Chinese is topic-prominent and uses constructions (*e.g*. ba, topicalisation) that diverge from the subject-prominent baseline implicit in many AI grammar-checking systems. Fourth, register and pragmatics are heavily encoded through formulaic expressions, honorifics, and culturally specific politeness moves, where AI fluent at the lexico-grammatical surface may still fail. These four demands together justify a language-specific synthesis rather than inferring Chinese-target evidence from English-target evidence.

The Three-Pillar Evidence Problem

Three existing systematic / scoping reviews of AI in language education (Chen *et al*. 2025; Anonymous 2025; Qiao 2025) have already documented that the SSCI evidence base is dominated by English-as-target-language studies and concentrates pedagogically on writing. On a first reading of this picture, one would expect the Chinese-as-target-language subset to be quantitatively sparse and that to be the central reform-relevant finding.

A formal PRISMA-ScR screen specific to Chinese as the learner's target language tells a more interesting story. After locking the inclusion rules to JCR 2023–2024 SSCI/SCIE Q1 or Q2 venues, the 2021–2026 window, and the strict AI definition adopted in Section 2.2, the database-plus-citation search identified 248 records (Web of Science Core Collection 33; Scopus 198; ERIC 13; four additional known-item records added through citation chasing). After deduplication 206 unique records remained; title/abstract screening retained 61 for full-text and venue review; the primary set then settles at 33 entries after full-text verification (see Tables **1** and **4**; see Section 2.4 for the flow, Figure **1**; the PRISMA-ScR 20-item checklist is filed as Appendix B). The remaining included evidence base presents not a scarcity problem but a three-pillar evidence problem.

Pillar 1: thematic imbalance. Within the included primary set, writing × AI feedback (7 entries; 21.2%) and teacher-facing / practitioner research (9 entries; 27.3%) are the largest clusters and together account for 16 of 33 entries (48.5%; see Figure **2**). Reading comprehension is represented by a single mediation-correlational study with no strict-AI intervention, and Chinese character / handwriting acquisition has zero Q1/Q2 entries despite the centrality of orthography to Chinese pedagogy. The K-12 cell is empty for direct learner interventions once a Q4 venue is excluded at the venue-verification step (see Section 2.4). The two field-level meta-analyses we identify (Hou and Min 2026; Lyu *et al*. 2025) each report substantial pooled effects for L2 speaking and L2 chatbot outcomes (Hedges' g = .61 and g = .608 respectively) but neither includes a single Chinese-as-target-language primary study among their 16+31 = 47 included investigations – an exact field-level mirror of our screen-level imbalance.

Pillar 2: design-maturity gap. Exactly one entry in the included primary set is a true randomised controlled trial: Q. Wang (2026) randomly assigned 80 beginning CFL learners to a 16-week 2× 2 factorial design crossing Interaction Partner (GAI vs. peer) with Interaction Structure (strong vs. weak). Four further entries qualify as longitudinal across one semester or more (a 1-year SQD professional-development trajectory, a 14-week multimodal-GenAI task-based course, a 12-week one-

group ChatGPT writing study, and a 6-week Sora vocabulary quasi-experiment). The remaining entries report sample sizes below one hundred or use within-subject quasi-experimental, qualitative, or cross-sectional structural-equation designs. We identify no replication, no pre-registered protocol, no public release of de-identified data, and – following the Chinese-EFL reclassification noted above – no latent-growth-curve analysis with L2 Chinese as target in the primary set. Nineteen of the 33 primary entries use samples below 100, so the modal empirical design remains small or single-site.

Pillar 3: ethics, originality, and cognitive dependence are under-examined. Five primary entries treat AI ethics, over-reliance, AI literacy, or AI-related threat appraisal as a primary outcome: Sun *et al*. (2026) on self-regulated learning phases and originality; Yan and Zhang (2025) on hindrance-appraisal mediation; Sun and Chan (2026) on the ABCE framework whose Ethical dimension is one of four pillars; Yijen Wang (2026) on critical AI literacy and authenticity as central themes; and Xia *et al*. (2024) on practitioner ethical caveats. A further set of entries treats ethics as a secondary theme rather than a primary outcome (see Section 3.10). We identify no validated CSL/CFL-specific AI-ethics framework in the screened evidence.

These three pillars define the problem this review synthesises and the agenda it proposes.

- **What This Review Adds**

We use AI-supported pedagogical innovation in an operational sense: a study is relevant when AI changes at least one component of the teaching-learning design – the feedback source or timing, the interaction partner, the representational mode, the teacher knowledge base, the assessment process, or the governance of learner autonomy. This definition distinguishes learner-intervention evidence from teacher-readiness and AI-literacy evidence; both are mapped, but they are not treated as equivalent evidence of learner gains.

The review is organised around three research questions:

RQ1: Which AI technologies, AI-related constructs, and pedagogical tasks are represented in Q1/Q2 Chinese-target evidence?

RQ2: What study designs, sample-size profiles, and quality signals characterise the included evidence?

RQ3: What task, learner-stage, and ethics gaps matter most for CSL/CFL pedagogical innovation and reform?

This review makes three contributions that, to our knowledge, no existing systematic or scoping review of AI in language education has combined. First, we provide the first synthesis explicitly restricted to Chinese-as-target-language learner populations in JCR 2023–2024 SSCI/SCIE Q1 or Q2 venues across 2021–2026, under a strict AI boundary definition that separates AI from pre-AI rule-based tools, pure multimodal stimuli, and pure virtual reality. The screen is documented at the level of database, query string, date, dedup, title/abstract decision, and venue-quartile verification (see Section 2 and the accompanying audit log in paper_rewriting_output/).

Second, we replace the “scarcity” framing that the rest of the field’s English-target reviews would predict for the Chinese-target subset. The included primary set holds 33 entries, not a handful, and the field is in a publication burst rather than a plateau. We therefore foreground the three-pillar evidence problem introduced in Section 1.3: thematic imbalance, design-maturity gap, and the under-examination of ethics and cognitive dependence. This frame is internally consistent with the empirical density visible in Figure **2** and with the quantitative reading developed in Section 3.10.

Third, we translate the three pillars into a prioritised research agenda R1–R7 (Section 5) that is anchored to specific primary entries as points of extension and to specific gaps as targets for new evidence. Each agenda item carries an explicit pillar tag so that future replications and design-based studies can be positioned against the same empirical map.

As artificial intelligence becomes mainstream in language education, research attention has shifted from technical exploration to practical teaching reform. Unlike general literature reviews that merely sort out existing studies, this scoping review takes teaching innovation and educational reform as the core orientation. By systematically sorting the quality empirical evidence in Q1/Q2 journals, identifying thematic imbalance, methodological defects and ethical risks in current research, this study not only maps the current development status of AI-assisted CSL/CFL teaching, but also provides evidence-based direction for classroom practice optimization, curriculum iteration and teaching evaluation reform. It bridges the gap between cutting-edge AI technology and front-line Chinese language teaching, and helps researchers and educators avoid blind technology application, so as to realize the deep integration of AI and Chinese language teaching rather

than superficial tool substitution. This is the core practical value of the present review for teaching innovation.

METHOD

Scoping-Review Protocol

We followed the PRISMA Extension for Scoping Reviews (Tricco *et al*. 2018). The manuscript is therefore a *scoping review*, not a systematic effectiveness review. A scoping rather than systematic review was chosen for three reasons. First, the field is heterogeneous in technology (large language models, automatic speech recognition, intelligent personal assistants, multimodal AI, educational robotics), in pedagogical task (writing, speaking, vocabulary, reading, teacher education), and in study design (qualitative, quasi-experimental, randomised controlled, latent profile, structural equation, longitudinal growth). Second, the outcome measures across the primary studies are non-comparable in their current form; pooling effect sizes is not defensible. Third, our aim is to map what has been studied and to identify where the field is empirically and methodologically uneven, which is a scoping-review aim rather than a meta-analytic one. We therefore *do not* promise effect-size pooling.

Inclusion and Exclusion Criteria

Inclusion

We retained records in two evidence categories. For the primary evidence set, a record had to satisfy all five criteria: (i) the venue is indexed in the JCR 2023-2024 release in SSCI or SCIE with a best-quartile of Q1 or Q2; this is a sampling boundary, not a proxy score for individual-study quality; (ii) the publication year is between 2021-01 and 2026-05; (iii) the learner's target language is Chinese (CSL, CFL, heritage Chinese, TCSOL, or international Chinese education); (iv) the manipulated or assessed variable meets the operational AI definition below; and (v) the article is empirical research published in English. Systematic or scoping reviews, technology comparators, and Chinese-EFL comparators were not counted in the primary set; when relevant, they were retained only as supporting evidence for field-level comparison.

Exclusion

We excluded a record if any of the following applied: (E1) the learner population is Chinese L1 mother-tongue "yuwen" / Chinese-language arts or native-speaker Chinese composition; (E2) Chinese students are the population but the target language is English (Chinese-EFL learners are kept only as a cross-language comparator in Section 4); (E3) the intervention does not meet our AI definition; (E4) the venue is not Q1 or Q2 in the JCR 2023–2024 SSCI/SCIE release; (E5) the publication is outside the 2021–2026 window; (E6) the document type is a conference paper, preprint, dissertation, editorial, or technical report; (E7) the dataset is a duplicate of a previously screened record; (E8) the full-text could not be retrieved through any of the access channels documented in Appendix D (full-text retrieval log).

AI definition (boundary rule)

Under the operational AI definition, we include: large language models and generative AI systems (*e.g*., ChatGPT, GPT-4, GPT-4o, Claude, DeepSeek); automatic speech recognition coupled to a learner model; intelligent tutoring systems; adaptive recommendation; chatbots and intelligent personal assistants with natural-language understanding; multimodal AI with model-level inference (*e.g*., DALL-E or Sora generation within an instructional loop); deep-learning–based assessment; and educational robotics with AI components. We treat the following as technology-enhanced comparators, not as AI evidence (Table **3**): rule-based acoustic visualisation tools such as Praat; pure multimodal stimuli without model-level inference; pure virtual reality without an AI component; traditional flashcard and computer-assisted instruction.

Two cross-disciplinary inclusions to be transparent about

Two venues require explicit disclosure. Acta Psychologica (Elsevier; JCR 2024 best-quartile Q1 in Arts & Humanities, Q2 in Developmental and Educational Psychology) is a psychology venue. Three primary entries in our set appear in this venue, and we adopt the best-quartile rule to retain them while flagging the cross-disciplinary character. International Journal of Adolescence and Youth (Taylor & Francis; JCR 2024 Q2 in Sociology) indexes one primary reading-comprehension study; sociology is a non-traditional venue for a CSL reading study, but the study itself measures reading comprehension and AI literacy, and is retained on those grounds.

One venue-driven exclusion at the venue-verification step

Chinese as a Second Language Research (De Gruyter; SJR Q4 in both Education and Linguistics) is excluded by criterion E4. The single study affected – which would

otherwise have been the only direct K-12 learner intervention in the primary set – is therefore reported only as background in Section 3.4 and in Appendix F, Table **F1**. This venue exclusion is the sole reason the K-12 learner cell in Figure **2** is empty after venue verification.

- **Search Strategy**

The search was completed on 2026-05-27. Three databases were queried as primary sources: Web of Science Core Collection (SCI-EXPANDED and SSCI editions, refined to Document Type = Article or Review, Language = English, with Web of Science Categories Education & Educational Research, Linguistics, Language and Linguistics, and Psychology Multidisciplinary); Scopus (Document Type = Article or Review, Language = English); and ERIC (Peer-reviewed = true, Publication Type = Journal Articles). The verbatim queries for each database, including the publication-year limiter *PY=(2021–2026)* for Web of Science and

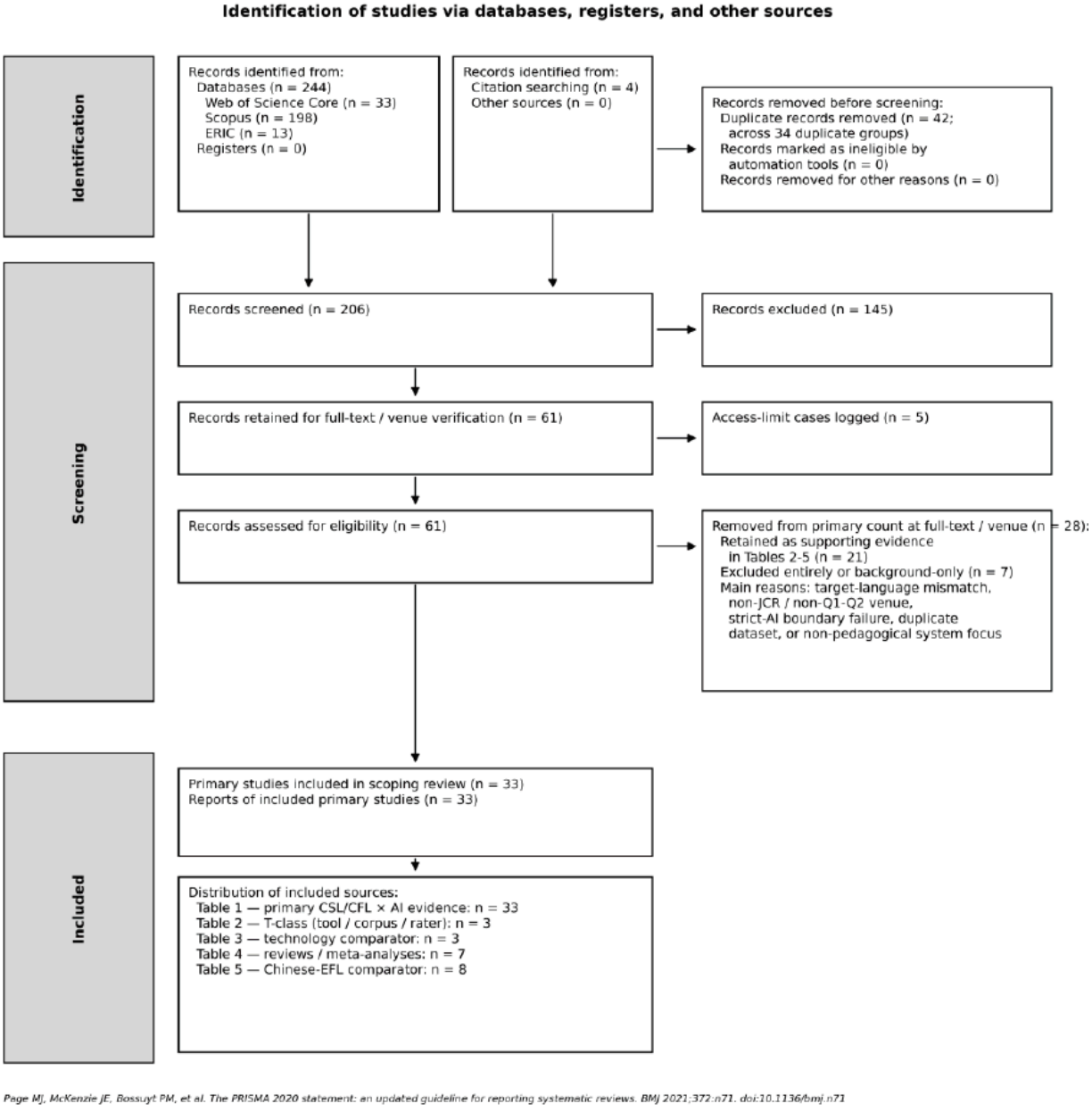

Figure 1: PRISMA-ScR flow diagram for the 2026-05-27 database-plus-citation search. Web of Science Core Collection 33; Scopus 198; ERIC 13; four additional known-item records via citation chasing (n = 248 total); 206 unique records after deduplication; 61 retained at title/abstract screening; 33 primary entries after full-text and venue verification, plus 3 T-class tool-evaluation entries (Table 2), 3 technology-enhanced comparators (Table 3), 7 reviews / meta-analyses (Table 4), and 8 Chinese-EFL cross-language comparators (Table 5).

PUBYEAR > 2020 AND PUBYEAR < 2027 for Scopus, are reproduced in the search log **paper_rewriting_output/prisma_log.md.**

JCR Q1/Q2 status is not embedded in the database query; it is verified post hoc against the JCR 2024 release for every retained record. The venue-verification log is reported as Appendix E; ten distinct venues required verification, seven met Q1/Q2 (six in Education or Educational Research; one in Sociology), and three failed the criterion (Appendix F, Table **F1**).

- ***Additional records identified through citation chasing***

Four primary records identified through backward citation chasing from three meta-analytic reviews in Table **4** were absent from the locked database export. Each of the four meets the strict inclusion criteria; the most likely reason for their absence from the database export is that the Web-of-Science refinement-by-category step and the Scopus title-abstract-keyword matching missed the phrasings 'intelligent personal assistant', 'AI-enhanced multimodal training', 'AI literacy', and 'non-Chinese students' in the absence of the broader acronym AI in the titles. Their inclusion is therefore a planned use of the PRISMA-ScR 'additional records identified through other sources' branch, not an ad-hoc add-back. Full details are recorded in Appendix C (screening log).

- **Screening and Selection**

Across the three databases plus the four additional records, 248 records were identified. Duplicate removal – by exact normalised DOI or by exact normalised title where DOIs were missing – reduced this to 206 records (42 duplicates removed across 34 duplicate groups). Title and abstract screening, conducted against the inclusion / exclusion taxonomy E1–E8, retained 61 records (52 clear includes plus 9 borderline retentions) for full-text and venue-quartile verification; 145 records were excluded at this step. The full counts and exclusion-reason tallies appear in Figure **1** and in Appendix C (screening log).

Full-text and venue-quartile verification separated the 61 retained records into a primary set and supporting evidence types. The primary set contains 33 Chinese-target empirical studies (Table **1**). Twenty-one further records are retained only as supporting evidence: three item-level AI tool / rater evaluations (Table **2**), three technology-enhanced comparators (Table **3**), seven systematic reviews or meta-analyses (Table **4**), and eight Chinese-EFL cross-language comparators (Table **5**). Seven records from the full-text / venue step were excluded from all mapped evidence categories or kept only as background notes because their venues were non-JCR / non-Q1-Q2, their AI claim failed the boundary rule, or their target language was not Chinese. One key reclassification is Zhu *et al*. (2025), whose 12-week latent-growth-curve study targets L2 English, not L2 Chinese; under criterion E2 it is moved to Table **5** (C8) rather than the primary set. The transparency log of every reclassification appears in Appendix F, Table **F1**.

- **Coding Scheme**

Each retained record was coded on eight dimensions: (1) AI technology (LLM/GenAI, ASR, ITS, adaptive recommendation, chatbot/IPA, multimodal AI, deep-learning assessment, robotics, multi-component); (2) pedagogical task (writing, speaking / pronunciation, listening, reading, vocabulary, characters / handwriting, pragmatics, assessment, teacher knowledge or professional development); (3) learner stage (K-12, higher-education, heritage, mixed); (4) study design (qualitative single-site, qualitative multi-site, quasi-experimental, randomised controlled, latent profile, structural equation, longitudinal growth, meta-analysis); (5) sample size class (<30, 30–99, 100–299, ≥ 300); (6) outcome class (learner gain, learner perception / engagement, teacher knowledge / efficacy, system output, hybrid); (7) theoretical frame (TAM, TRAM, TPACK, self-determination theory, self-regulated learning, flow theory, none stated); and (8) ethics treatment (primary outcome, secondary observation, not addressed). The codebook operationalising each dimension is reproduced in the Appendix.

- **Quality Appraisal**

Quality appraisal is reported as a descriptive design appraisal consistent with scoping-review practice. JCR Q1/Q2 status is used only to delimit the sampling frame; it is not treated as evidence that an individual study is high quality. We did not assign MMAT/JBI scores or present a formal risk-of-bias matrix. Instead, each entry was described using design type, sample-size class, comparator or repeated-measure status, triangulation, missing-data or attrition reporting where available, and explicit ethics reporting. These signals are used in the design-maturity synthesis rather than as a single numeric quality score. We do not pool effect sizes; where individual effect sizes are reported in the original studies, we cite them as point estimates only. The design-maturity analysis in Section 3.10 relies on the design-class and sample-size-class codes rather than on any pooled estimate.

RESULTS

Overall Scope of the Evidence Base

The PRISMA-ScR screen described in Section 2.3 and Section 2.4 (search date 2026-05-27) returns a mapped evidence base of five distinct evidence types. The primary set is 33 Chinese-as-target-language empirical studies meeting the strict AI boundary definition in JCR 2023–2024 SSCI/SCIE Q1 or Q2 venues (Table **1**); three further studies report item-level AI tool or rater evaluations (Table **2**, summarised in Section 3.3); three technology-enhanced comparator studies (Table **3**, Praat-based feedback, multimodal input, virtual-reality vocabulary) provide a non-AI baseline for the same pedagogical tasks; seven systematic reviews or meta-analyses (Table **4**) provide the cross-language field baseline; and eight Chinese-EFL comparator studies (Table **5**) are reserved for the cross-language discussion in Section 4. Four of the 33 primary entries entered the evidence base through the PRISMA-ScR "additional records identified through other sources" branch documented in Section 2.3. The next eight subsections walk the primary set cluster by cluster; Section 3.10 closes with the three-pillar reading of the whole base.

Technology × Task Evidence Grid

Figure **2** maps the 33 primary entries onto a technology–task grid. The first rows represent AI intervention technologies; the final rows separate AI-related constructs, study designs, and ethics/practitioner themes so that technology, design, and theme are not conflated. Three features of the grid are load-bearing for the rest of Section 3. First, the densest cells cluster around writing and teacher-facing evidence. Second, two cells are explicitly shaded as empty gaps: G2 (Chinese characters and handwriting, an entire column with zero Q1/Q2 entries) and G3 (Reading, where only a single mediation-correlational entry exists, with no strict-AI intervention). Third, the K-12 marker dagger attaches only to the teacher-side primary-school lesson-planning entry; no direct K-12 learner intervention survives the venue-verification step (see Section 2.2). The cluster subsections (Section 3.3–Section 3.9) read each non-empty cell against its evidence; Section 3.10 reads the imbalance, maturity, and ethics features of the grid as a whole.

Writing × AI Feedback

The writing cluster is the densest in the primary set (7/33, 21.2%). Across comparator, outcome, acceptance, and self-regulated-learning studies, the evidence supports a cautious claim: LLM feedback can improve holistic writing products and confidence, but the mechanism is uneven across proficiency levels and higher-order learning constructs. GPT-4 and ChatGPT-assisted conditions

Figure 2: Technology × task evidence grid. Rows distinguish intervention technologies from construct/design/theme evidence; cells list Table 1 entry IDs. Empty cells are empirical gaps.

generally outperform no-interaction or traditional-comparison conditions on writing-product indicators, while native-speaker, teacher, and peer feedback remain stronger anchors for relational, cultural, or developmental judgement. The cluster therefore supports AI feedback as a scalable writing scaffold, not as a replacement for human mediation. Item-level tool and rater evaluations are kept outside the primary learner-outcome count and reported separately in Table **2**; the English-target LGCA comparator is retained in Table **5** rather than in the primary set.

- **Speaking, Pronunciation, Dialogue**

The speaking, pronunciation, and dialogue cluster contains five primary entries (15.2%) and supplies the strongest design anchor in the whole review: one 16-week randomised 2 x 2 trial. The cluster separates two affordances. First, AI can function as a dialogue partner that increases interaction opportunities through IPA, chatbot, or voice interfaces. Second, ASR-linked systems can function as pronunciation feedback engines for tone and intelligibility. The evidence is promising but bounded: interaction quantity and structured practice improve, and the RCT shows large effects for interaction partner and interaction structure, yet tonal accuracy, spontaneous fluency, and interactional naturalness remain sensitive to task design and human participation. The Praat comparator in Table **3** and the dialogue-CALL meta-analysis in Table **4** provide useful baselines, but neither supplies direct Chinese-target meta-analytic evidence.

- **Vocabulary and Multimodal Generative AI**

The vocabulary and multimodal-GenAI cluster contains three primary entries (9.1%). Together they show that multimodal generation can support vocabulary learning, task products, motivation, and lesson-design activity, especially when image or video generation is embedded in a structured instructional loop. The strongest evidence is a six-week quasi-experiment using Sora-supported vocabulary instruction; the longest intervention is a 14-week task-based multimodal-GenAI course. Both point to a familiar boundary condition: gains are clearer for product quality, motivation, and vocabulary-related outcomes than for spontaneous oral fluency, tonal accuracy, or stable individual-interest measures. The K-12-adjacent evidence remains teacher-side planning rather than direct learner intervention, and the absence of Chinese character or handwriting acquisition evidence remains a central empty cell.

- **Reading and AI Literacy**

The reading cluster is the thinnest cluster in the primary set: one Q1/Q2 entry (3.0%), and it is mediation-correlational rather than interventional. The study links AI literacy to L2 Chinese reading comprehension through orthographic and morphological knowledge among Form-4 ethnic-minority learners in Hong Kong. This makes the entry valuable for locating AI literacy inside Chinese reading development, but it does not license causal claims about AI-mediated reading instruction. The review therefore treats AI-scaffolded reading comprehension as an open intervention gap.

- **Learner Profiles, Beliefs, AI Literacy**

The learner-profile and engagement cluster contains five entries (15.2%) spanning latent profiles, SEM models, and one quasi-experimental motivation study. The cluster shows that AI-related learner experience is not a single attitude variable: engagement, grit, resilience, anxiety, enjoyment, autonomy, and usage frequency form distinct profiles and pathways. The strongest intervention-like evidence is a ChatGPT-supported writing study with large immediate and delayed task-motivation effects. The remaining profile and SEM studies are useful for explaining heterogeneity but cannot establish causal transitions between learner profiles. One included engagement study remains a declared population-label boundary case, so its generalisation to L2-Chinese learners should be read cautiously.

- **Teacher TPACK, Robotics, Professional Development**

The teacher-side cluster is the second-densest cluster in the primary set (9/33, 27.3%). It is dominated by framework, scale, readiness, profile, and professional-development evidence rather than direct classroom implementation trials. Across AI-TPACK, AI readiness, AI-flow, ISTE-adapted competencies, robotics, and GenAI professional development, three signals converge: teacher self-efficacy and AI attitudes mediate readiness; structured exposure and PD shift teachers toward higher-order integration; and risks around dependence, creativity loss, infrastructure, and teacher-student interaction remain persistent. The cluster is therefore strong for construct formation and teacher-facing reform implications, but weaker for observed classroom enactment and student-learning outcomes.

Pedagogical-Model Integration – a Framework Gap

The primary set contains framework-adjacent studies but no empirically tested CSL/CFL pedagogical model that sequences AI, teacher, and peer roles across a complete learning loop. Existing studies validate readiness, flow, and AI-TPACK constructs or nominate scaffolding components such as cognitive support, load management, trust calibration, and oral-output bridging. They do not yet specify how these components should be weighted across task preparation, AI interaction, human feedback, revision, assessment, and reflection. This is the framework gap targeted by R7.

Three Concurrent Problems: Imbalance, Maturity, and Ethics

This subsection is the analytical centre of the review. Taking Figure **2** together with Table **1**, we read the included primary evidence base of 33 entries as three concurrent problems rather than as a single scarcity problem.

Pillar 1 – thematic imbalance

Within the 33 primary entries, writing x AI feedback (7/33, 21.2%) and teacher-facing / practitioner evidence (9/33, 27.3%) are the two largest clusters and together account for 16 of 33 entries (48.5%). Engagement and learner beliefs contribute five entries (15.2%), speaking / pronunciation five (15.2%), vocabulary / multimodal work three (9.1%), project-based scaffolding two (6.1%), reading one (3.0%), and game-based tutoring one (3.0%). Chinese character / handwriting acquisition and direct K-12 learner interventions remain empty after venue verification. The same imbalance appears in the field-level comparison: two L2 speaking/chatbot meta-analyses together aggregate 47 primary studies, none of which has Chinese as the target language.

Pillar 2 – design-maturity gap

Design maturity is uneven. The primary set contains exactly one true RCT (P16), one additional quasi-experimental study with random assignment (P37), four repeated or longitudinal intervention entries of six weeks or more, and one further repeated-measures pronunciation study without a control group. Nineteen primary entries (57.6%) report samples below 100, and many studies remain single-site, cross-sectional, qualitative, or SEM/profile-analytic. We identify no replication, no pre-registered protocol, no public de-identified dataset, and no latent-growth-curve analysis of L2 Chinese. The field has reached critical mass, but the modal design is still not built for cumulative inference.

Pillar 3 – ethics, originality, and cognitive dependence are under-examined

Ethics, originality, and cognitive dependence are visible but not yet consolidated. Five primary entries treat ethics, originality, AI literacy, threat appraisal, or dependence as primary outcomes; a further ten treat over-reliance, trust, integrity, privacy, or creativity loss as secondary themes. Across these studies, the recurrent concern is not simply that learners use AI, but that shallow use may substitute for metacognitive engagement, original production, or teacher-student dialogue. However, the constructs remain fragmented: there is no validated CSL/CFL-specific AI-ethics framework, no replicated over-reliance measure, and no shared originality or cognitive-dependence instrument.

Two gaps now closed in part, three still open

Two gaps are partly closed and three remain open. K-12 CSL/CFL evidence is partly addressed by teacher-side planning and a secondary-school correlational reading study, but direct K-12 learner intervention remains empty. The RCT / longitudinal gap is partly addressed by one RCT and several repeated-intervention entries, but replication and open data remain absent. Reading intervention, Chinese character / handwriting acquisition, and consolidated ethics-as-primary-outcome instruments remain fully open priorities for R1-R7.

Effect-size summary across clusters (non-pooled)

We do not pool effect sizes, but the reported point estimates indicate that effects range from small to large depending on task and design. Writing studies report within-subject gains around d = 0.71-0.91; speaking/pronunciation studies include large multivariate partial eta-squared values in the RCT and significant ASR-training interactions; multimodal vocabulary studies report partial eta-squared around .10-.144; engagement and teacher-side SEM studies report meaningful but design-limited path coefficients. The field-level meta-analytic baselines are positive (B5 g = .61; B6 g = .608) but contain no Chinese-target primary studies, so they cannot substitute for CSL/CFL cumulative evidence.

DISCUSSION

Innovation Lever: Feedback Latency and Scaffolding Granularity

Across the 33 selected empirical studies, the most prominent pedagogical innovation brought by AI is not technological iteration, but the restructuring of two core teaching elements: feedback latency (the time interval between learners' output and targeted feedback) and

Table 1: Condensed evidence map for primary CSL/CFL x AI empirical studies (N = 33).

Cluster	Primary IDs	n (%)	Main evidence signal	Remaining gap
Writing x AI feedback	P1, P2, P4, P5, P9-P11	7 (21.2%)	LLM feedback, assessment, acceptance, and writing-affect evidence.	Multi-site RCTs and proficiency-stratified evidence remain limited.
Speaking / pronunciation / dialogue	P13-P16, P32	5 (15.2%)	IPA, ASR, voice-GenAI, chatbot, and one 16-week RCT.	K-12 longitudinal CAPT evidence remains absent.
Vocabulary / multimodal GenAI	P18-P20	3 (9.1%)	Multimodal task-based learning, lesson planning, and Sora-supported vocabulary.	Character and handwriting acquisition remain empty cells.
Reading / AI literacy	P21	1 (3.0%)	AI-literacy mediation evidence for L2 Chinese reading.	No strict-AI reading intervention is identified.
Project-based scaffolding	P22-P23	2 (6.1%)	GenAI-supported PBL and speaking-readiness evidence.	Shared pedagogical model and replication remain missing.
Teacher / practitioner evidence	P24-P31, P39	9 (27.3%)	AI-TPACK, readiness, flow, profiles, PD, robotics, and teacher ethics evidence.	Field deployment and observed classroom integration remain under-tested.
Learner profiles / engagement	P33-P37	5 (15.2%)	Profile, affect, grit, resilience, motivation, and engagement evidence.	Causal profile-transition evidence remains limited.
Game-based tutoring	P42	1 (3.0%)	LLM tutor evidence with accuracy and satisfaction-gain disconnect.	Replication and learning-process evidence are needed.
Total primary empirical evidence	P1, P2, P4, P5, P9-P11, P13-P16, P18-P37, P39, P42	33 (100%)	Primary CSL/CFL x AI empirical evidence base.	Evidence is clustered, not evenly distributed across tasks and learner stages.

Note: Percentages are calculated against the primary empirical set (N = 33). Supporting evidence is separated from the primary set: P3, P6, and P7 report item-level or system-output evidence and appear in Table 2; P8 is reclassified to Table 5 because its target language is L2 English. The compact map preserves the primary IDs used in Figure 2 and in the cluster-level synthesis.

Table 2: T-class evidence: item-level / corpus / rater evaluations of AI tools (not learner-outcome studies).

ID	Authors (Year)	Journal	AI tech	Key system-level finding (verbatim)
P3	Wu & Lin (2026)	IJAL Q1	ChatGPT-4o	200 texts (40 original + 40 expert + 120 GPT in 3 prompts). Lex MANOVA Wilks' $\Lambda = 0.371$, $p < .001$, $\eta^2_p = 0.281$. GPT subject-density > expert ($\eta^2_p = 0.107$); GPT idiom density < expert ($\eta^2_p = 0.326$).
P6	Lu, Liles, Ma (2025)	Ass. Wri. Q1	ChatGPT-4.5 + DeepSeek-V3	33 learners × 198 essays + 8 raters under MFRM. ChatGPT severity -1.46 logits ≈ experienced teachers; DeepSeek +0.72 ≈ novices; no genre bias after Bonferroni.
P7	Zhang, Xu, Nguyen (2025)	IJAL Q1	ChatGPT-4	25 Vietnamese HS CSL + 1,104 sentences. Revision rates ANOVA $F(2, 825) = 112.22$, $p < .001$; ChatGPT M = 6.26, Senior M = 1.69; ChatGPT accuracy 5.25% vs. Senior 47.01%.

Table 3: Technology-enhanced comparator entries (non-AI pre-AI tools that were initially nominated and reclassified at the AI-definition step).

ID	Tool / Configuration	Reason for comparator (not primary) classification
C-tech1	Praat-based tone-feedback	Rule-based acoustic visualisation; no AI inference layer. Retained as historical reference for ASR-driven feedback (Section 3.4).
C-tech2	Multimodal stimuli only	Multimodal input without model-level inference. Retained as historical reference for vocabulary cluster (Section 3.5).
C-tech3	Virtual reality vocabulary	VR without an AI component. Retained as comparator for the vocabulary cluster.

scaffolding granularity(the refinement degree and hierarchical design of instructional support). Three robust empirical findings support this conclusion.

First, Chen and Gong (2025) verified that AI writing feedback brought substantial learning gains ($t = 6.53$, $p < 0.001$, partial $\eta^2 = 0.317$). In traditional classrooms,

Table 4: Systematic reviews and meta-analyses (k studies of L2/foreign-language AI; Chinese-target coverage flagged for each).

ID	Authors (Year)	Journal	Scope, pooled effect, Chinese-target inclusion
B1	Chen *et al*. (2025)	JCAL	AI-enabled assessment in language learning 2012–2024 (k ≈ 25); Chinese sub-sample documented at the cited level.
B2	Zhang *et al*. (2025)	ILT	GenAI scoping review in language education 2022–2024 (≈ 43 SSCI); proportion English-target large by margin.
B3	Qiao *et al*. (2025)	IJAL	AI in language education systematic review; broad scope; cited at abstract level.
B4	Yang *et al*. (2025)	JECR	AI assessment review; cited at abstract level.
B5	Hou & Min (2026)	ReCALL	Dialogue-CALL three-level meta. k = 16, n = 89 ES, g = .61 [.34, .89], Q = 329.75. Zero Chinese-target primary study.
B6	Lyu, Lai, Guo (2025)	IJAL	Chatbot meta-analysis. k = 31, n = 2,943, g = .608 [.43, .79], I^2 = 80.75%. Zero Chinese-target (28 EFL + 1 EN/FR + 1 ES + 1 KFL).
B7	Şahin Kızıl *et al*. (2025)	JCAL	Chatbot systematic review 2020–May 2024, k = 33 (narrative); 4 of 33 are Chinese-target (Chen 2020, Divekar 2021, Wu 2024, Li/Li/Cho 2024).

Table 5: Chinese-EFL cross-language comparator entries (Chinese L1 learners with English as target).

ID	Citation hint	Role in Section 4.4
C1	Crompton *et al*. BJET review of AI in EFL	Latency-collapse mechanism (English target)
C2	AI-IDLE CALL study	Chinese learners of English; informal digital learning
C3	EAIT IDLE listening study	EFL listening with AI
C4	EJE classroom rapport & AI-literacy	Cross-language attitude evidence
C5	EJE AI-literacy outcomes	Cross-language outcome evidence
C6	CALL/AIALL transition framework	Pedagogical-model comparator (English)
C7	Cross-linguistic RCT (EFL)	Practice-loop validation in EFL
C8	Zhu, Wang, Qin (2025) ILT 12-week LGCA	181 Chinese-L1 undergraduates preparing NPEE-EII English; AF slope advantage 1.85, p < .001; reclassified from primary at full-text step

teachers usually deliver written feedback days after learners finish drafts; by contrast, generative AI such as ChatGPT provides instant feedback, enabling learners to revise wording, sentence structure and rhetorical logic multiple times in a single learning session. Second, Zhou (2026) found that instant AI guidance significantly boosted learners' task motivation (d ≈ 1.03 at immediate post-test, d ≈ 0.70 at delayed post-test). When learners receive timely inspiration during idea generation, they gain stronger autonomy and participation willingness, which optimises the affective dimension of language learning. Third, the meta-analysis conducted by Hou and Min (2026) (overall Hedges' g = 0.61 for L2 speaking) proved that semantic constraints and interactive modes − the key embodiments of scaffolding granularity − exert greater influence on learning outcomes than learners' proficiency level or teaching cycle.

To sum up, AI fundamentally changes the traditional teaching logic: it moves feedback from the end-of-lesson summary correction to real-time micro-support during learning tasks. This core mechanism explains why AI can improve teaching efficiency and learning experience in CSL/CFL education. Meanwhile, existing studies also expose inherent boundaries of this mechanism, covering practical application risks and ethical challenges.

This innovative model still faces two critical constraints in real teaching scenarios. First, Yijen Wang (2026) proposed the illusion of preparedness. Although real-time micro-feedback effectively optimises learners' written performance, including vocabulary range, grammatical accuracy and task completion quality, it cannot improve core oral abilities such as tonal accuracy and spontaneous fluency. This proves that AI feedback has differentiated effects on different language skills, and its advantages are more concentrated on written production rather than impromptu oral expression. Second, J. Zhang *et al*. (2025) revealed the trust paradox: excessive trust and reliance on AI tools are negatively correlated with learners' speaking proficiency (β = -0.104). Excessive

pursuit of instant feedback will gradually weaken learners' independent thinking and language practice ability, resulting in cognitive dependence.

In practical teaching, therefore, the key to applying AI lies in reasonable calibration: make full use of low-latency feedback to improve learning efficiency, while setting clear boundaries to maintain learners' critical engagement and independent creation. The following sections further elaborate on concrete classroom application schemes, theoretical basis, as well as corresponding norms for ethical AI use.

- **Practical Classroom & Curriculum Implementation Cases**

Drawing on the AI tools and research findings presented in this review, targeted operational frameworks are proposed to integrate artificial intelligence into in-class instruction, curriculum design and assessment for CSL/CFL education.

For writing teaching, a hybrid model combining generative AI and human guidance is applicable. ChatGPT and GPT-4o may function as real-time revision tools. Learners complete drafts independently before receiving instant feedback on vocabulary, grammar and sentence structure from AI systems. Instructional focus can then shift to discourse logic, cultural expression and rhetorical style, which cuts down repetitive correction work and enhances teaching efficiency. For pronunciation and tone training, ASR-based AI systems work effectively in oral classrooms. Such systems capture pronunciation in real time, display pitch contour diagrams visually and deliver targeted corrections for Mandarin tones. This form of one-to-many personalised training is difficult to achieve within conventional teaching environments. In vocabulary instruction, multimodal AI such as Sora and DALL-E generates pictures and short videos corresponding to target Chinese words. Immersive situational contexts are therefore created to consolidate vocabulary knowledge and raise learning engagement.

Traditional single-lesson curriculum units can be restructured into iterative practice loops, as noted in previous sections. Undergraduate CFL programmes may adopt a full learning framework featuring AI-assisted brainstorming prior to tasks, AI-based interactive practice throughout activities, and peer as well as teacher evaluation upon completion. The framework can be simplified for primary and secondary Chinese courses, where short-cycle AI interactive activities fit the attention characteristics of young learners. Integration of AI literacy training into regular coursework also supports appropriate and standardised use of relevant technologies among both learners and teaching staff.

Artificial intelligence serves well as a supplementary assessment instrument. It undertakes quantitative evaluation including vocabulary accuracy checks, grammatical error identification and text length analysis. Qualitative evaluation concerning content creativity, pragmatic appropriateness and oral fluency remains the responsibility of teachers. This collaborative mode brings together the operational efficiency of machine-based assessment and the interpretive depth of human evaluation.

- **Boundary Conditions and Teaching Governance**

The evidence also shows where AI-supported innovation should not be over-claimed. Learner-intervention studies and teacher-readiness studies answer different questions: the former estimate what happens to learner performance or motivation under an AI-supported task, while the latter estimate whether teachers are prepared to design or govern such tasks. The strongest practice implication is therefore not "use AI", but calibrate AI to learner stage, task type, and human support. Low-proficiency writing learners can experience negative emotion when corrections exceed their comprehension capacity; oral video tasks can improve vocabulary range and task achievement without improving tone, fluency, or delivery confidence; and high trust in AI can coexist with weaker speaking outcomes. These boundary conditions are especially important for Chinese because tone, character structure, register, and pragmatics are not interchangeable with English-target feedback problems.

For classroom use, the governance implication is concrete. AI-supported CSL/CFL tasks should require process evidence (prompt/output logs, revision histories, or learner explanations), keep teacher or peer feedback in the loop for pragmatics and spontaneous oral performance, and make AI-use disclosure part of assessment design rather than an afterthought. Character learning, handwriting feedback, reading scaffolding, and direct K-12 learner interventions remain empirical blank cells, so recommendations for those areas should be treated as design hypotheses rather than evidence-based prescriptions.

- **Guidelines for Responsible AI Use and Academic Integrity for Educators**

Practical guidance regarding AI ethics, academic integrity and the prevention of cognitive dependence can be

established for teaching practice. Clear specifications for AI application can be set at the start of courses to define appropriate scenarios such as idea brainstorming and vocabulary inquiry, while restricting improper use including direct copying of AI-generated content and complete reliance on artificial intelligence to finish assignments.

Evaluation systems can attach importance to process-based evidence covering AI interaction records, revision drafts and oral statements, rather than focusing merely on final outputs. This approach helps mitigate excessive reliance on technological tools among learners.

In writing and oral language courses, comparisons between personal expressions and AI-generated content can be arranged to examine differences in vocabulary, pragmatics and cultural connotations, which facilitates independent creation. Artificial intelligence is positioned as a supplementary educational tool rather than a replacement for individual thinking.

Classroom discussions regarding AI inaccuracies, content homogenisation and excessive tool use help foster rational attitudes and uphold academic integrity in Chinese language learning.

- **Cross-language Comparator and Structural Reading of Sparsity**

The Chinese-EFL comparator entries in Table **5** (C1–C8) confirm the latency-collapse and granularity-shift mechanisms operate on English-target outcomes too, with the practice-loop curriculum unit operationalising through structured AI-supported task cycles. Three qualifications restrict the import to CSL/CFL: tones and sentence-final particles impose constraints on AI-assisted oral practice that English-target studies do not test; register and pragmatics in Chinese (the *ba*-construction, topicalisation, politeness moves) remain weak spots in current LLM output; and handwriting/character acquisition has no Chinese-EFL analogue.

The sparsity of Chinese-target evidence relative to the size of the CSL/CFL learner population reflects three structural causes rather than a research-output deficit. First, much CSL/CFL educational- technology research is published in Chinese in CSSCI core journals that do not appear in JCR 2023–2024 SSCI/SCIE indexes and are therefore systematically excluded from any Q1/Q2-anchored review. Second, the SSCI-indexed Chinese-target empirical record is concentrated in a few research networks (Wenzhou/UiTM/HK in P22/P23/P29; Sichuan/Lincoln in P24/P26/P27; Hebei/HIT in P25/P31), each building density along a single argument line. Third, the post-2022 GenAI inflection arrived later in CSL/CFL field-level uptake than in English-target research; the 2025–2026 publication burst visible in the primary set is the field's first post-inflection wave and is still in publication transit. The three pillars are therefore a snapshot of a field in active burst rather than a permanent feature, and R1–R7 is calibrated to that snapshot.

The sparsity of Chinese-target evidence relative to the size of the CSL/CFL learner population reflects three structural causes rather than a research-output deficit. Much CSL/CFL educational-technology research appears in Chinese-language CSSCI core journals, which are not included in the JCR 2023‑2024 SSCI/SCIE indexes and thus excluded from this Q1/Q2-focused review. SSCI-indexed empirical studies targeting Chinese learners tend to cluster within limited research networks, with each group concentrating on a single line of inquiry. The widespread adoption of generative AI within CSL/CFL research also occurred later compared with English language education. The noticeable growth of relevant publications from 2025 to 2026 marks the first wave of research following the popularization of generative AI, and related studies are still being produced. The three core research issues identified in this review represent a current state of a rapidly developing field instead of fixed traits, and the proposed R1‑R7 research agenda is formulated based on this status.

AI-assisted language teaching shares universal features across alphabetic languages such as English and Chinese. Artificial intelligence generally enhances feedback efficiency, diversifies interactive approaches and optimises assessment mechanisms, and the effect of reduced feedback latency applies to foreign language instruction in general. Meanwhile, distinct traits distinguish AI implementation for Chinese education. As a tonal language, Mandarin requires dedicated AI speech recognition and visual tone feedback functions that are unnecessary for non-tonal languages. The logographic nature of Chinese characters also sets higher technical standards for AI-powered handwriting recognition and stroke analysis, which differ fundamentally from tools designed for alphabetic writing systems. Additionally, topic-prominent structures, the ba construction and sophisticated polite expressions create challenges for AI grammatical checking and pragmatic analysis, leading to lower accuracy compared with English language applications. Research priorities also vary across languages. While AI-related studies on English learning centre on writing and speaking abilities, Chinese language education places extra emphasis on character

acquisition and tonal practice, forming a distinct research landscape. Practices developed for other languages cannot be directly applied to CSL/CFL contexts, and language-specific design is essential for AI integration into Chinese teaching.

- **A Prioritised Research Agenda (R1–R7)**

Each agenda item below is anchored to the three pillars introduced in Section 1.3 and Section 3.10. We mark the pillar(s) each item targets, name the primary entry that the item extends (where applicable), and indicate the design class that the field would most usefully add.

- ***R1 – Multi-site, pre-registered RCTs of GenAI corrective feedback in CSL writing across proficiency levels***

(*Pillars 1&2*). The writing cluster is dense yet contains no L2-Chinese RCT or LGCA. A multi-site RCT randomising proficiency-stratified learners to teacher-only, GenAI-only, and teacher–GenAI hybrid conditions, pre-registered with open data and a small-to-medium effect power calculation, would close the design-maturity gap. Writing is currently the most mature research cluster (21.2% of all studies) with abundant preliminary findings. Prioritizing this agenda is because writing is the core skill of CSL/CFL, and multi-site pre-registered RCTs can solve the defects of small sample and single-site research in existing literature. It can generate high-evidence conclusions to guide the large-scale promotion of AI writing feedback, so it is ranked first.

- ***R2 – Semester-or-longer ASR/CAPT trials for Mandarin tones at the K-12 stage***

(*Pillars 1&2*). Direct K-12 learner interventions are absent. The closest evidence – a single-session adult tone-training study – shows AI feedback can move production but does not test the K-12 age range where motivational and attentional constraints differ. A semester-long field trial in real K-12 classrooms with documented teacher–AI integration and transparent attrition reporting would close both gaps. Direct K-12 learner intervention is a key blank in current research. Primary and secondary education is the main stage for beginners to establish Chinese tone perception. Long-term classroom trials can fill the research gap for young learners and expand the application scope of AI tone training, so it is listed as the second priority.

- ***R3 – AI-scaffolded reading comprehension at the intermediate-to-advanced level***

(*Pillar 1*). The reading cluster holds one mediation-correlational study and zero strict-AI interventions. Closing this requires designs in which AI is embedded in the reading task (*e.g.*, LLM-mediated text simplification with reader-controlled granularity, generative scaffolding of inferential questions) and outcomes are measured against a non-AI counterpart. Reading comprehension is the core input skill of language learning. Supplementing AI reading intervention research can balance the unbalanced theme distribution of current studies, which is an urgent supplementary direction for the field.

- ***R4 – Chinese character and handwriting acquisition with AI under a design-based-research programme***

(*Pillar 1*). The empty cell. Handwriting recognition with stroke feedback, generative stroke decomposition, and character-model-aware spaced repetition are technologically available but absent from the Q1/Q2 empirical record. A multi-iteration DBR programme with a defined acquisition outcome would establish a first anchor. Chinese character and handwriting learning is the most distinctive difficulty of Chinese language teaching, yet there is zero high-quality Q1/Q2 research. This agenda targets the most prominent blank field of CSL/CFL AI teaching, with strong linguistic and teaching pertinence.

- ***R5 – AI-TPACK curricula and field deployment for pre-service TCSL teachers***

(*Pillars 1&2*). The teacher cluster is large but the modal study is single-site cross-sectional SEM. Next step: a multi-cohort study embedding an AI-TPACK curriculum in a TCSL programme, tracking trainees into field placements with both self-reported AI-TPACK and observed classroom integration as outcomes. Teacher-related studies account for the largest proportion (27.3%), but most stay at theoretical exploration. Promoting curriculum design and on-the-job application can realize the transformation from teacher research to classroom practice, supporting the sustainable development of AI teaching.

- ***R6 – Ethics, originality, and cognitive dependence as primary outcomes***

(*Pillar 3*). Five primary entries treat ethics/over-reliance as primary outcomes but use distinct constructs. The field needs shared instruments: a cross-study over-reliance measure, an originality assessment that handles Chinese rhetorical norms, and an equity evaluation across heritage versus non-heritage learners. Ethical risks and learners' cognitive dependence are common hidden dangers of AI application. Unified measurement tools and evaluation frameworks can standardize the research

system of AI ethics in CSL/CFL, and provide normative guidance for the whole field.

- ***R7 – CSL/CFL-specific pedagogical-model frameworks grounded in empirical data.***

Pillars 1 & 3. The framework studies presently in the primary set are scale-development studies (an AI-flow framework, an AI-readiness scale). What the field does not have is a pedagogical-model framework – specifying how AI, teacher, and peer roles are sequenced and weighted at each stage of a CSL/CFL learning loop – that is grounded in empirical implementation rather than in conceptual review. The Acta Psychologica CFL project-based-learning study and the System AI-flow study together suggest the components; an empirical synthesis study that operationalises a CSL/CFL-specific pedagogical-model framework and tests it in two or more sites would close G7. Based on all the above empirical research results, constructing an exclusive pedagogical model is the final goal of teaching innovation. It can integrate all research conclusions into a complete teaching system, so it is set as the long-term priority agenda.

LIMITATIONS

Five limitations bound the present synthesis. First, the review is a scoping rather than meta-analytic synthesis; with N = 33 heterogeneous primary entries spanning eight clusters, multiple AI technology / construct classes, and mixed design classes, effect-size pooling is not defensible, and we do not promise it. Readers seeking pooled estimates should consult the two field-level meta-analyses verified in the present screen (Hou and Min 2026; Lyu *et al*. 2025) as baselines, with the caveat that neither includes any Chinese-target primary study (0 of 47 combined included investigations).

Second, the Q1/Q2 venue cap on JCR 2023–2024 SSCI/SCIE indexes excludes ESCI-only venues such as Computers & Education: Artificial Intelligence (whose JCR indexing changed during the review window) and several local-Chinese-language CSSCI venues catalogued in Appendix F, Table **F2**. Several relevant Chinese-target studies are therefore visible only as background sources in the present synthesis even though they contribute substantively to the field's Chinese-language discourse. This venue restriction is a deliberate design choice for Q1/Q2-anchored comparability, not an evaluation of the excluded work.

Third, grey literature, conference papers, preprints, dissertations, and editorial pieces were excluded by criterion E6. Some fast-developing AI-pedagogy threads – in particular, voice-mode generative AI for K-12 Mandarin learners and adaptive character recognition under deep-learning models – are visible in the conference and preprint record but not yet in the JCR 2024 SSCI/SCIE journal record, and the present synthesis under-represents them as a consequence.

Fourth, two cross-disciplinary venue inclusions warrant disclosure as limitations rather than as primary findings. Three primary entries appear in Acta Psychologica (Elsevier; JCR 2024 best-quartile Q1 in Arts & Humanities, Q2 in Developmental and Educational Psychology); one primary entry appears in the International Journal of Adolescence and Youth (T&F; JCR 2024 Q2 in Sociology). The retention of these entries follows the best-quartile rule but introduces cross-disciplinary heterogeneity into the included primary set that a strict best-fit-category review would not contain. We have not re-weighted these entries downward, but readers should note the cross-disciplinary inclusion when comparing density across clusters.

Fifth, three boundary calls are documented here in the interest of transparency. (i) Population-label boundary and Acta Psychologica concentration. Chen (2025a) is framed in its title, keywords, and discussion as a study of "Chinese language learners" within "International Chinese Education," but the Methods section describes a generic sample of 382 Xuzhou-region university students "with a range of academic majors" without an explicit L1 or CFL marker. We retain the entry on the strength of the journal-keyword and discussion framing but mark it as a boundary case; readers should treat its generalisation to L2-Chinese learners specifically as provisional. A related sensitivity check addresses the cross-disciplinary concentration in Acta Psychologica (3 of 33 primary entries: P22, P23, P35): if all three Acta Psy entries are excluded, the primary set reduces to N = 30 with cluster densities shifting to 23.3% writing, 30.0% teacher-facing / practitioner evidence, and 13.3% engagement – preserving the concentration pattern in Pillar 1 and the design-maturity gap of Pillar 2 without substantive change. The three-pillar conclusions are therefore robust to the Acta Psy retention decision. (ii) Within-paper inconsistency. J. Zhang *et al*. (2025) reports the Trust-in-AI predictor of speaking proficiency as a statistically significant negative path in its narrative but as $\beta = -0.104$, $p = .074$ (n.s. at $\alpha = .05$) in the source regression-output table; the paper also reports a single descriptive standard deviation as both 0.81 (source descriptive-statistics table) and 4.81 (narrative). We code the entry on its verbatim source-table values, but the within-paper inconsistency

means the trust–proficiency direction should be treated as a directional rather than a confirmatory claim. (iii) Target-language reclassification. The 12-week latent-growth-curve study by Zhu *et al*. (2025) was retained at venue verification but reclassified at the full-text step to Table **5** (C8) because the target language is L2 English, not L2 Chinese. This decision is logged transparently and removes the only LGCA from the primary set.

CONCLUSION

This scoping review of N = 33 Chinese-as-target-language empirical studies in JCR 2023–2024 SSCI/SCIE Q1 or Q2 venues (2021–2026) re-frames the Chinese-target AI-in-language-education evidence base as a three-pillar problem rather than a scarcity problem. **Pillar 1 (thematic imbalance)**: writing × AI feedback (7/33, 21.2%) and teacher-facing / practitioner evidence (9/33, 27.3%) are the two largest clusters and together hold 16 of 33 entries (48.5%); reading is represented by a single mediation-correlational study; Chinese characters and handwriting are empty (G2); direct K-12 *learner* interventions are empty after venue verification (G1); and two field-level meta-analyses jointly aggregate 47 L2 speaking and L2 chatbot primary studies of which zero have Chinese as the target language. **Pillar 2 (design-maturity gap)**: exactly one randomised controlled trial (Q. Wang 2026), four further repeated or longitudinal intervention entries of six weeks or more, no replication, no pre-registered protocol, no latent-growth-curve analysis of L2 Chinese, and no public de-identified data. **Pillar 3 (ethics under-examined)**: five primary entries treat AI ethics, over-reliance, AI literacy, or threat appraisal as primary outcomes but use distinct constructs and instruments; no validated CSL/CFL-specific AI-ethics framework exists. The **R1–R7** agenda translates the three pillars into prioritised designs – pre-registered RCTs (R1), K-12 ASR/CAPT trials (R3), AI-scaffolded reading (R3), characters under DBR (R4), AI-TPACK implementation (R5), ethics-as-primary-outcome instruments (R6), and CSL/CFL pedagogical-model frameworks (R7) – as the field's next empirical step. A Tripartite Interactive Framework for AI-assisted Innovative Chinese Language Teaching defines the roles and interaction logic of artificial intelligence, teachers and learners within integrated teaching systems. Artificial intelligence functions as an efficient auxiliary tool, undertaking repetitive tasks such as real-time feedback, error detection, personalised training and learning resource recommendation to streamline teaching workflows. Teachers serve as core leaders and quality controllers, taking charge of curriculum design, task arrangement, pragmatic guidance, ethical supervision and overall evaluation throughout the learning process. Learners occupy the central position of learning activities, engaging in autonomous study and interactive practice via AI tools while following instructional guidance to accomplish knowledge acquisition and ability development. Components within this framework form an interconnected and complementary system. Technology exists to support instructional practice, and all teaching activities centre on learner development, which represents the core direction of AI-enabled innovation in CSL and CFL education.

APPENDIX A. CODING SCHEME

This appendix operationalises the eight coding dimensions named in Section 2.5 into allowable codes and one-sentence definitions. The codebook was used for every retained record in Table **1**, Tables **2**-**5**, and Appendix F.

D1 – AI Technology

Single dominant technology of the manipulated or assessed variable, drawn from the following closed set: **LLM/GenAI** (large language model or generative AI as the intervention; *e.g*., ChatGPT, GPT-4, GPT-4o, Claude, DeepSeek); **ASR** (automatic speech recognition coupled to a learner model); **ITS** (intelligent tutoring system with a learner-state model); **Adaptive-Rec** (adaptive recommendation of content); **Chatbot/IPA** (chatbot or intelligent personal assistant with natural-language understanding); **Multimodal-AI** (multimodal generation or recognition with model-level inference, *e.g*., DALL-E, Sora, voice-mode GenAI); **DL-Assess** (deep-learning-based assessment); **Robotics** (educational robotics with AI components); **Multi** (more than one of the above, with no single dominant technology). A record is assigned exactly one code; multi-technology studies receive **Multi**.

D2 – Pedagogical Task

Primary task addressed by the study, drawn from: **Writing**, **Speaking-Pron** (speaking or pronunciation), **Listening**, **Reading**, **Vocabulary**, **Characters** (Chinese characters / handwriting), **Pragmatics** (pragmatic or cultural-rhetorical

task), **Assessment** (the task is assessing learner work), **Teacher-K** (teacher knowledge or professional development). A record may be tagged with a secondary task only if the secondary task is explicitly measured.

- **D3 – Learner Stage**

K-12, **HE** (higher education), **Heritage** (heritage CFL), **Mixed**. Where the participants are pre-service teachers planning K-12 lessons rather than K-12 students using AI, the stage is **HE** with a teacher-planning note in the audit log.

- **D4 – Study Design**

Qual-Single (qualitative single-site), **Qual-Multi** (qualitative multi-site), **QuasiExp** (quasi-experimental; non-randomised intervention with comparison), **RCT** (*randomised controlled trial*; defined here as random allocation of participants to two or more conditions *plus* a pre-registered or method-section-stated allocation procedure and a defined control or comparison condition). Quasi-experimental designs with random assignment but without a pre-stated allocation protocol – *e.g*. Zhou (2026), who used random allocation to a peer-collaborative comparator but did not pre-register the protocol – are coded as **QuasiExp**, not **RCT**. **LPA** (latent profile analysis), **SEM** (structural equation model including mediation or moderation), **LGCA** (latent growth curve analysis), **Meta** (meta-analysis or systematic review).

- **D5 – Sample Size Class**

S (N < 30), **M** (30–99), **L** (100–299), **XL** (N ≥ 300). For meta-analyses and systematic reviews we record k (number of included studies) rather than N.

- **D6 – Outcome Class**

Gain (learner outcome, *e.g*., proficiency, accuracy, fluency), **Perception** (learner perception, motivation, engagement), **Teacher-Knowledge** (teacher TPACK, self-efficacy, AI readiness), **System** (system or tool output, item-level reliability or accuracy), **Hybrid** (combines two or more of the above).

- **D7 – Theoretical Frame**

TAM, **TRAM**, **TPACK**, **SDT** (self-determination theory), **SRL** (self-regulated learning), **Flow**, **Other** (any other named theoretical frame), **None** (no explicit theoretical frame).

- **D8 – Ethics Treatment**

Primary (ethics, over-reliance, or originality is a primary outcome), **Secondary** (ethics or over-reliance is addressed but as a secondary theme), **Incidental** (mentioned only in Limitations or Discussion), **None** (not addressed).

APPENDIX B. PRISMA-SCR CHECKLIST

This appendix maps the manuscript to PRISMA-ScR reporting items. It is a reporting checklist, not a protocol registration record.

Table B1: PRISMA-ScR checklist.

PRISMA-ScR item	How addressed	Manuscript location
Title/Abstract/Rationale/Objectives	Title and abstract identify a PRISMA-ScR scoping review of AI in Chinese as a target language; RQ1-RQ3 specify the review objectives.	Title, Abstract, Introduction
Eligibility criteria and information sources	JCR Q1/Q2 boundary, year window, Chinese-target condition, operational AI definition, document type, and databases are reported.	Method 2.2-2.3
Search and selection	Web of Science, Scopus, ERIC, citation chasing, deduplication, title/abstract screening, full-text and venue verification are reported.	Method 2.3-2.4; Figure 1
Data charting and data items	Eight coding dimensions are specified and operationalised in Appendix A.	Method 2.5; Appendix A

Critical appraisal	The manuscript uses descriptive design appraisal rather than formal risk-of-bias scoring.	Method 2.6; Results 3.10
Synthesis of results	Evidence is synthesised by cluster, technology-task map, and three-pillar framework.	Results, Discussion, R1-R7 agenda
Limitations and conclusions	Scoping-design, venue-boundary, grey-literature, cross-disciplinary, and boundary-call limitations are stated.	Limitations and Conclusion

APPENDIX C. SCREENING LOG

This appendix reports the record-flow counts used in Figure **1** and the evidence-category split used in Tables **1**-**5**.

Table C1: Screening log.

Stage	Count	Notes
Web of Science Core Collection	33 raw records	Database source
Scopus	198 raw records	Database source
ERIC	13 raw records	Database source
Backward citation chasing	4 additional records	Other source
Total identified	248 records	Before deduplication
After deduplication	206 unique records	42 duplicates removed
Title/abstract retained	61 records	52 clear includes + 9 borderline retentions
Primary empirical set	33 records	Chinese-target empirical studies in Q1/Q2 SSCI/SCIE venues
Supporting evidence	21 records	3 T-class, 3 technology comparators, 7 reviews/meta-analyses, 8 Chinese-EFL comparators
Excluded/background-only after full-text/venue step	7 records	Venue, AI-boundary, or target-language reasons

APPENDIX D. FULL-TEXT RETRIEVAL LOG

This appendix summarises the full-text retrieval status used for criterion E8. It reports final evidence categories rather than provisional draft IDs.

Table D1: Full-text retrieval log.

Evidence category	Final status	Use in manuscript
Primary empirical evidence	33 records	Full-text and venue verification completed for the primary set
T-class tool/rater evidence	3 records	Retained outside primary learner-outcome count
Technology comparators	3 records	Retained as non-AI historical or design comparators
Reviews/meta-analyses	7 records	Retained as field-level baseline evidence
Chinese-EFL comparators	8 records	Retained for cross-language comparison, not primary CSL/CFL evidence
Unavailable full text	Excluded under E8 where unresolved	No unavailable full text is counted in the primary set

APPENDIX E. VENUE-VERIFICATION LOG

Venue status was checked after database retrieval because JCR quartile is a sampling boundary, not a database-query term.

Table E1: Venue-verification log.

Venue	Verification decision	Manuscript consequence
Education Sciences	Q1/Q2 JCR-indexed education venue	P9, P42 retained
Asia-Pacific Education Researcher	Q1/Q2 JCR-indexed education venue	P20 retained
Professional Development in Education	Q1/Q2 JCR-indexed education venue	P29 retained
AERA Open	Q1/Q2 JCR-indexed education venue	P37 retained
Asia Pacific Journal of Education	Q1/Q2 JCR-indexed education venue	P30 retained
Educational Technology & Society	Q1/Q2 JCR-indexed education venue	P32 retained
Acta Psychologica	Best-quartile retained with cross-disciplinary caveat	P22, P23, P35 retained and sensitivity checked
International Journal of Adolescence and Youth	Q2 Sociology; retained with caveat	P21 retained
Chinese as a Second Language Research	Failed Q1/Q2 JCR criterion	Direct K-12 learner study background only / Appendix F, Table F1
Future in Educational Research	No JCR impact factor assigned in 2024 release	Background only / Appendix F, Table F1
Applied Corpus Linguistics	No qualifying JCR quartile assignment in review frame	Background only / Appendix F, Table F1

APPENDIX F. RECLASSIFICATION AND LOCAL-CONTEXT TABLES

Table F1: Reclassified-out / background-only records

Original ID	Final treatment	Reason
P8	Table 5, C8	Target language is L2 English, not L2 Chinese.
P17	Background only	Venue fails the Q1/Q2 criterion.
P40	Background only	Future in Educational Research had no JCR Impact Factor assigned in the 2024 release.
P41	Background only	Applied Corpus Linguistics is Scopus-indexed without a JCR quartile assignment.
P43	Background only	Venue not Q1/Q2 SSCI/SCIE.
P44	Background only	Technical / system-oriented venue outside the Q1/Q2 education-linguistics frame.
C-tech4	Background only	ESCI / Q3 SJR venue; retained only as a CAPT background comparator.

Table F2: Local policy / Chinese-language context sources

ID	Source type	Role in the review
S2-1	PRC Ministry of Education guidance	Context for generative AI governance in K-12 education.
S2-2	International Chinese Language Education journal discussion	Chinese-language field context for GenAI in international Chinese education.
S2-3	Chinese education-technology policy discourse	Context for AI-empowered education and the strong-education-nation agenda.
S2-4	CNKI / CSSCI bibliometric context	Background signal for Chinese-language scholarship excluded by the JCR Q1/Q2 sampling boundary.

REFERENCES

[1] Anonymous. 2025. “A Scoping Review of Empirical Studies on Generative Artificial Intelligence in Language Education.” Innovation in Language Learning and Teaching, ahead of print. https://doi.org/10.1080/17501229.2025.2509759.

[2] Chen, Chen, and Yang Gong. 2025. “The Role of AI-Assisted Learning in Academic Writing: A Mixed-Methods Study on Chinese as a Second Language Students.” Education Sciences 15 (2): 141. https://doi.org/10.3390/educsci15020141.

[3] Chen, Jingjing. 2025a. “Examining the Role of Chinese Language Learners’ Grit and Self-Efficacy on Their Engagement in Artificial Intelligence-Driven Settings.” Acta Psychologica 259: 105357. https://doi.org/10.1016/j.actpsy.2025.105357.

[4] Chen, Jingjing. 2025b. “Modeling Chinese Second Language Learners’ Motivation, Engagement, and Resilience in AI-

Enhanced Contexts: A Self-Determination Theory." Learning and Motivation 92: 102199. https://doi.org/10.1016/j.lmot.2025.102199.

[5] Chen, Xiaodong, Xiaosheng Zhou, and Ying Soon Goh. 2026. "How Chinese as a Foreign Language Learners Use Generative AI for Oral Script-Writing: A Qualitative Perspective on Cognitive Scaffolding in Project-Based Learning." Acta Psychologica 262: 106071. https://doi.org/10.1016/j.actpsy.2025.106071.

[6] Chen, Xieling et al. 2025. "A Systematic Review and Meta-Analysis of AI-Enabled Assessment in Language Learning: Design, Implementation, and Effectiveness." Journal of Computer Assisted Learning, ahead of print. https://doi.org/10.1111/jcal.13064.

[7] Cheng, Bing, Kun Liao, Yang Xiang, Yipeng Zou, Xiaolin Zhang, and Yang Zhang. 2025. "Development and Validation of an AI-Enhanced Multimodal Training Program: Evidence from Non-Native Mandarin Tone Learning." Computer Assisted Language Learning, ahead of print. https://doi.org/10.1080/09588221.2025.2571696.

[8] Dong, Xuesong, and Hong Wang. 2026. "Incorporating Robotics and Artificial Intelligence (AI) into Teaching Chinese as a Second Language in Higher Education: Unveiling the Potential Challenges and Opportunities." European Journal of Education 61 (1): e70442. https://doi.org/10.1111/ejed.70442.

[9] Fang, Lu, Ge Tang, and Lu Zhang. 2025. "Helpful or Harmful? Comparative Study of Perceived and Actual Effectiveness of LLM-Driven Tutors in Game-Based CFL Learning." Education Sciences 15 (11): 1502. https://doi.org/10.3390/educsci15111502.

[10] Gao, Anna, Aiqing Yu, Guanyao Xu, Guy Trainin, and Xiaoyang Luo. 2026. "Developing an AI-Assisted Learning Flow Framework for Pre-Service Teachers of Chinese as a Second Language: Investigating the Influence of Teacher Educators' Support on AI-Flow and Self-Efficacy in Teaching." System, ahead of print. https://doi.org/10.1016/j.system.2026.104025.

[11] Hou, Zhuohan, and Shangchao Min. 2026. "Dialogue-Based Computer-Assisted Language Learning Systems for Second Language Speaking Development: A Three-Level Meta-Analysis." ReCALL 38 (1): 40–56. https://doi.org/10.1017/S0958344025100268.

[12] Kızıl, Aysel Şahin, Blanka Klimova, Marcel Pikhart, and Antigoni Parmaxi. 2025. "A Systematic Review of the Recent Research on the Usefulness of Chatbots for Language Education." Journal of Computer Assisted Learning 41 (2): e70001. https://doi.org/10.1111/jcal.70001

[13] Lan, Yu-Ju, Scott Grant, and Hui-Chin Yeh. 2025. "The Use of Virtual Chatbots to Support Chinese as a Foreign Language Learners' Communication Skills Through Scaffolded Self-Directed Learning." Educational Technology & Society 28 (2): 434–52. https://doi.org/10.30191/ETS.202504_28(2).SP04.

[14] Li, Nuoen, and Yu Liang. 2025. "Teachers' AI Readiness in Chinese as a Foreign Language Education: Scale Development and Validation." System, ahead of print. https://doi.org/10.1016/j.system.2025.103597.

[15] Liu, Shu-Hsiang Johnny. 2026. "Effects of Voice-Based ChatGPT on Chinese-as-a-Second-Language Learners' Mandarin Pronunciation." International Journal of Applied Linguistics, ahead of print. https://doi.org/10.1111/ijal.70213.

[16] Lu, Yixuan, Xinhui Liles, and Xiaoyan Ma. 2025. "GenAI and Human Assessments of L2 Chinese Writing: Interrater Reliability and Rater Bias." Assessing Writing 66: 100989. https://doi.org/10.1016/j.asw.2025.100989.

[17] Lyu, Boning, Chun Lai, and Jingyi Guo. 2025. "Effectiveness of Chatbots in Improving Language Learning: A Meta-Analysis of Comparative Studies." International Journal of Applied Linguistics 35 (2): 834–51. https://doi.org/10.1111/ijal.12668.

[18] Qiao, Anonymous. 2025. "Artificial Intelligence for Language Learning: A Systematic Review of Its Design, Theoretical Foundations, Implementation, and Impact." International Journal of Applied Linguistics, ahead of print. https://doi.org/10.1111/ijal.70034.

[19] Shan, Zhaoyang, Zhangyuan Song, Xu Jiang, Wen Chen, and Luyao Chen. 2025. "Complementing but Not Replacing: Comparing the Impacts of GPT-4 and Native-Speaker Interaction on Chinese L2 Writing Outcomes." Behavioral Sciences 15 (4): 540. https://doi.org/10.3390/bs15040540.

[20] Sun, Jingyue, Yu Wang, and Zhihui Qian. 2025. "Examining the Influence of Individual-Level Cultural Values on CFL Learners' Acceptance of ChatGPT for Chinese Learning." Interactive Learning Environments 33 (5): 3393–407. https://doi.org/10.1080/10494820.2024.2443785.

[21] Sun, Lan, Fan Jin, Kening Zhou, Wing Mui Cheung, and Chin-Hsi Lin. 2026. "From Clueless to Confident: How ChatGPT Transforms Academic Writing in Chinese as a Second Language." International Journal of Applied Linguistics 36 (2): 1235–51. https://doi.org/10.1111/ijal.12849.

[22] Sun, Zheng Liang, and Yau Yu Chan. 2026. "The Role of AI Literacy in Chinese L2 Reading Comprehension: Evidence from Adolescent Learners in Hong Kong." International Journal of Adolescence and Youth 31 (1). https://doi.org/10.1080/02673843.2025.2610107.

[23] Tricco, Andrea C., Erin Lillie, Wasifa Zarin, et al. 2018. "PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation." Annals of Internal Medicine 169 (7): 467–73. https://doi.org/10.7326/M18-0850.

[24] Wang, Qian. 2026. "The Effects of Generative AI, Peer Interaction, and Interaction Structure on Oral Performance and Perceived Feedback Among Beginning Learners of Chinese as a Foreign Language." International Journal of Applied Linguistics, ahead of print. https://doi.org/10.1111/ijal.70186.

[25] Wang, Yijen. 2026. "Exploring Student Engagement with Multimodal Generative AI in Task-Based Chinese Language Learning." Computer Assisted Language Learning, ahead of print. https://doi.org/10.1080/09588221.2026.2645428.

[26] Wang, Yongliang. 2026. "Exploring the Impact of AI-Enhanced Language Tools on Multilingual Learners' Grit, Enjoyment, Anxiety, and Emotional Disengagement: A Positive Psychology 2.0 Perspective." International Journal of Multilingualism, ahead of print. https://doi.org/10.1080/14790718.2026.2641065.

[27] Wei, Wei, Yue Xu, and Zhimin Wang. 2026. "CSL Learners' Acceptance and Use of ChatGPT: An Extended Technology Readiness and Technology Acceptance Model." Humanities and Social Sciences Communications 13: 154. https://doi.org/10.1057/s41599-025-06452-w.

[28] Wu, Juan, Yuxin Li, Jianrong Zhou, and Shiya Chen. 2024. "The Impact of Intelligent Personal Assistants on Mandarin Second Language Learners: Interaction Process, Acquisition of Listening and Speaking Ability." Computer Assisted Language Learning, ahead of print. https://doi.org/10.1080/09588221.2024.2317849.

[29] Wu, Junjie, and Yaping Lin. 2026. "ChatGPT for Text Simplification in Chinese as a Second Language Textbooks: A Comparative Study with Expert Adaptations." International Journal of Applied Linguistics, ahead of print. https://doi.org/10.1111/ijal.70147.

[30] Xia, Jingfang, Yao Ge, Zijun Shen, and Mudasir Rahman Najar. 2024. "The Auxiliary Role of Artificial Intelligence Applications in Mitigating the Linguistic, Psychological, and Educational Challenges of Teaching and Learning Chinese Language by Non-Chinese Students." International Review of Research in Open and Distributed Learning 25 (3). https://doi.org/10.19173/irrodl.v25i3.7680.

[31] Xu, Guanyao, Aiqing Yu, Anna Gao, and Guy Trainin. 2025. "Developing an AI-TPACK Framework: Exploring the Mediating Role of AI Attitudes in Pre-Service TCSL Teachers' Self-Efficacy and AI-TPACK." Education and Information Technologies 30: 22471–95. https://doi.org/10.1007/s10639-025-13630-5.

[32] Xu, Guanyao, Aiqing Yu, Cong Xu, Xianquan Liu, and Guy Trainin. 2025. "Investigating Pre-Service TCSL Teachers' Technology Integration Competency Through a Content-Based AI-Inclusive Framework." Education and Information Technologies 30 (4): 4349–80. https://doi.org/10.1007/s10639-024-12982-8.

[33] Yan, Yaping, and Jie Zhang. 2025. "Exploring the Interplay of Academic Resilience, Cognitive Appraisals of GenAI, and Academic Engagement Among Pre-Service Chinese as a Foreign Language Teachers." European Journal of Education 60 (3): e70202. https://doi.org/10.1111/ejed.70202.

[34] Yao, Yuyu, Yiqun Zhu, Pan Li, Lin Zhang, and Xinmin Zhu. 2026. "The Application of Multimodal GenAI in Lesson Planning for Primary School L2 Chinese Vocabulary Teaching: A TPACK Perspective." Computer Assisted Language Learning, ahead of print. https://doi.org/10.1080/09588221.2025.2592010.

[35] Zang, Xuan, Nuoen Li, and Yu Liang. 2026. "Profiles of Chinese as a Foreign Language Teachers' Perceptions on Generative Artificial Intelligence: Antecedents and Outcomes." Asia Pacific Journal of Education, ahead of print. https://doi.org/10.1080/02188791.2026.2613806.

[36] Zhai, Xiuwen, and Wu Chen. 2025. "Unpacking Beliefs and Engagement in AI-Assisted Chinese Learning: A Latent Profile Analysis of Malaysian Chinese as a Foreign Language Learners." European Journal of Education 60 (2): e70101. https://doi.org/10.1111/ejed.70101.

[37] Zhang, Jie, Xiaosheng Zhou, and Ying Soon Goh. 2025. "Unpacking AI-Supported Chinese as a Foreign Language Learning: How Beginner-Level Learners' Cognitive and Motivational Factors Predict Speaking Proficiency." Acta Psychologica, ahead of print. https://doi.org/10.1016/j.actpsy.2025.105703.

[38] Zhang, Liang, Jingyu Xu, and Hong Anh Nguyen. 2025. "Linguistic Analyses of Written Corrective Feedback for Chinese as a Second Language: ChatGPT Versus Human Teachers." International Journal of Applied Linguistics, ahead of print. https://doi.org/10.1111/ijal.70053.

[39] Zhao, Xian, Danping Wang, and Guangxiang Leon Liu. 2026. "Innovating Chinese Vocabulary Learning Through Multimodal GenAI: The Motivational, Interest, and Attitudinal Shifts Among CSL Learners." The Asia-Pacific Education Researcher, ahead of print. https://doi.org/10.1007/s40299-026-01098-x.

[40] Zhao, Xinyi, and Danping Wang. 2026. "The Impact of ChatGPT's Feedback on L2 Chinese Learners' Writing Outcome, Confidence, and Emotions: A Mixed-Method Quasi-Experimental Study." Assessing Writing 68: 101027. https://doi.org/10.1016/j.asw.2026.101027.

[41] Zhou, Keyi, Hongyun Deng, Hong Ching Chan, and Chin-Hsi Lin. 2025. "Integrating Generative AI into Language Teachers' Professional Development: A Longitudinal Study Using the Synthesis of Qualitative Data (SQD) Model." Professional Development in Education, ahead of print. https://doi.org/10.1080/19415257.2025.2586643.

[42] Zhou, Xiaosheng. 2026. "Investigating L2 Learners' Task Motivation in Mandarin Chinese Essay Writing: The Role of ChatGPT in AI-Mediated Learning." AERA Open 12. https://doi.org/10.1177/23328584261435952.

[43] Zhou, Xiaosheng, Hongbin Wu, and Ying Soon Goh. 2026. "Evaluating ChatGPT-4o as an AI Assessor in Chinese as a Second Language Writing: Reliability Through Generalizability Theory, Feedback Actionability, and Teacher-Student Perceptions." Journal of Second Language Writing 73: 101311. https://doi.org/10.1016/j.jslw.2026.101311.

[44] Zhu, Ruoyi, Huizhen Wang, and Xiaobin Qin. 2025. "Longitudinal Comparison of AI, Exemplar, and Teacher Feedback for Sustainable L2 Writing Development: A Latent Growth Curve Analysis." Innovation in Language Learning and Teaching, ahead of print. https://doi.org/10.1080/17501229.2025.2586142.