Discover Computing

REVIEW Open Access

# A meta-analysis of generative AI effects on language proficiency and affective–cognitive outcomes in language learning

Mirka Saarela[1*], Sachini Gunasekara[1] and Prabha Kumarage[1]

*Correspondence:
Mirka Saarela
mirka.saarela@jyu.fi
[1]Faculty of Information Technology, University of Jyväskylä, Mattilanniemi 2, 40100 Jyvaskyla, Finland

**Abstract**
Generative Artificial Intelligence (GenAI) tools are increasingly integrated into language education, yet their overall effectiveness and the factors shaping their impact remain unclear. This meta-analysis synthesizes evidence from 51 studies reporting 175 independent effect sizes to evaluate the impact of GenAI tools on second and foreign language learning outcomes. Results indicate that GenAI tools produce large, statistically significant effects on language proficiency and affective–cognitive outcomes. Moderator analyses revealed that effects vary by intervention duration, instructional setting, learner characteristics, and target language. Stronger effects were generally observed in informal learning environments, for productive skills, and in studies addressing less commonly taught languages. These findings support sociocultural and interactionist theories of second language acquisition and highlight the need for tailored, context-sensitive integration of GenAI tools. The review offers practical guidance for educators, designers, and policymakers, and underscores the importance of continued research across diverse contexts and learner populations.

**Keywords** Meta analysis, Language education, Generative artificial intelligence, ChatGPT, Second and foreign language learning

## 1 Introduction

The recent rise of Generative Artificial Intelligence (GenAI) tools such as ChatGPT has sparked widespread interest in their potential applications within language education [1, 2]. Unlike earlier rule-based or task-specific systems, GenAI enables dynamic natural language generation, context-sensitive dialogue, and creative content production. These affordances make it more interactive, flexible, and adaptive than traditional AI systems, opening new possibilities for second and foreign language learning [3, 4]. Beyond language learning, meta-analytic evidence suggests that GenAI can improve learning outcomes more broadly, including performance and higher-order thinking across various educational domains [5].

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Given these affordances, it is unsurprising that GenAI has been rapidly adopted, especially since the release of ChatGPT in late 2022, in both formal classrooms and informal learning environments, with students and educators using these tools for writing assistance, speaking practice, and personalized feedback [6–11]. Earlier chatbot-based interventions showed promise in reducing learner anxiety and fostering low-stakes practice [12, 13], but were limited by scripted interactions and superficial language support. Recent advances in GenAI address many of these shortcomings, enabling richer and more personalized learner interactions [7, 11].

Despite these promising developments, the theoretical mechanisms underlying GenAI's effectiveness in language learning remain underexplored. Two prominent frameworks in second language acquisition–sociocultural theory and interactionist theory–offer valuable lenses for understanding how GenAI may support learning. From a sociocultural perspective, language development is mediated through tools and social interaction; GenAI tools can serve as digital mediators that scaffold learner performance within their zone of proximal development [14, 15]. At the same time, interactionist approaches emphasize the role of comprehensible input, negotiation of meaning, and corrective feedback in language acquisition [16]–affordances that GenAI tools can provide through real-time, adaptive dialogue. These perspectives suggest that GenAI may not only supplement instruction but actively shape learning processes when integrated in pedagogically sound ways.

Meta-analyses can quantify the effects of AI-based tools on language learning outcomes and provide empirical evidence for their potential benefits. For example, Lyu et al. [17] found moderate effects of chatbot use, Lee [18] reported improvements in writing and vocabulary through AI translation systems, and Xiu-Yi [7] synthesized AI interventions more broadly, observing moderate to large effects depending on learner age, instructional setting, and intervention duration. To date, only one meta-analysis [8] has specifically examined GenAI use in language learning, focusing on informal English learning and reporting small to medium effects, while highlighting the lack of research in structured educational contexts. Together, these findings underscore a critical gap: no meta-analysis has systematically evaluated GenAI's impact on second and foreign language learning across both formal and informal instructional contexts.

To address this gap, the present study synthesizes 51 experimental studies focusing on second and foreign language learning, published between 2023 and 2025, representing the first wave of research following ChatGPT's release in late 2022, and encompassing 175 independent effect sizes. The analysis addresses the following research questions:

- **RQ1**: What is the effect of GenAI tools on second and foreign language learning outcomes compared to other instructional approaches?
- **RQ2**: What learner, intervention, and contextual factors moderate the effects of GenAI tools on second and foreign language learning outcomes?

By aggregating diverse empirical findings and systematically exploring key moderators, this study provides the first meta-analytic synthesis of GenAI's effectiveness in second and foreign language education. It aims to offer a robust evidence base to guide educators, developers, and policymakers in effectively leveraging GenAI tools across diverse language learning contexts.

## 2 Literature review

This section reviews the existing literature on AI and GenAI applications in language education, focusing on previous meta-analyses, empirical findings, and relevant moderators. First, we examine prior meta-analytic studies that synthesize evidence on various AI tools and their educational effects (Section 2.1). Next, we discuss emerging research on the use of GenAI in formal and informal language learning settings, highlighting both opportunities and challenges (Section 2.2). Finally, we outline the key moderator variables identified in the literature and adopted in this meta-analysis to explain variability in GenAI intervention outcomes (Section 2.3).

### 2.1 Prior meta-analyses on AI in language education

Several meta-analyses have examined AI applications in language education. Lyu et al. [17] investigated the effects of chatbots on second and foreign language learning, synthesizing 31 studies and reporting a medium overall effect size ($g = 0.608$). Their analysis identified modality, mobility, the involvement of GenAI, and the type of comparison group as significant moderators. Lee [18] conducted a meta-analysis on machine learning-based translation tools, finding positive effects on writing accuracy and vocabulary development, with variations depending on task type and learner proficiency. Xu et al. [19] examined the relationship between intelligent tutoring systems and reading comprehension in L2 contexts, reporting modest but significant benefits, particularly when combined with teacher support. More recently, Xiu-Yi [7] conducted a comprehensive meta-analysis of AI interventions in L2 learning, analyzing 49 studies (79 reports) and reporting a moderate to large overall effect size (SMD = 0.982). That study identified significant moderating effects for learning environment, intervention duration, educational level, learner age, interactivity types, and feedback sources, with particularly strong effects for online and blended environments, medium-duration interventions, young adult learners, and productive skills such as speaking and writing. AI interventions also positively influenced learner motivation, anxiety reduction, and willingness to communicate, though effects on engagement, satisfaction, and self-regulation were more mixed [7].

A recent meta-analysis by Guan et al. [8] focused specifically on GenAI in informal, out-of-class English learning contexts. Synthesizing 15 experimental studies, the review reported small to medium positive effects on English proficiency ($g = 0.447$) and self-regulation ($g = 0.508$), with stronger effects for speaking skills ($g = 0.794$), but no significant impact on learning motivation. This study highlights both the promise of GenAI for autonomous language learning and the limited availability of empirical research in this area. A relevant contribution from outside the language domain is the meta-analysis by Wang and Fan [5], which analyzed 51 experimental studies published between November 2022 and February 2025. Their study assessed the impact of ChatGPT on students' overall learning performance, learning perception, and higher-order thinking, reporting a large positive effect on learning performance ($g = 0.867$) and moderately positive effects on learning perception ($g = 0.456$) and higher-order thinking ($g = 0.457$). Although not confined to language learning, the study underscores the broader educational potential of GenAI tools such as ChatGPT.

Although these prior meta-analyses have advanced our understanding of AI-based educational interventions, they predominantly focus on pre-GenAI technologies Lyu et

al. [17–19] or informal language learning [8]. Few syntheses have systematically evaluated the effects of GenAI tools, and none have addressed second and foreign language learning across both formal and informal settings. This gap motivates the present meta-analysis.

### 2.2 Generative AI in language education

GenAI builds upon decades of research in educational technology, including early developments in chatbots [20], automated feedback tools [21, 22], and intelligent tutoring systems [23]. Although these early tools focused on specific tasks like writing evaluation or grammar correction, recent advances in Large Language Models (LLMs) have enabled more interactive, dynamic, and context-aware learning experiences. Due to this shift, GenAI is now positioned as an effective tool for language learning that is capable of simulating conversation, promoting learner autonomy on an unprecedented scale, and providing personalized on-demand feedback [24–26].

Traditional methods of second language learning often follow fixed structures and standardized content, which may overlook the diverse needs and learning preferences of individual students. These approaches, while organized, can lack the flexibility to support varied learning styles or provide meaningful, interactive engagement. As a result, some learners may fall behind or struggle to meet expected outcomes [27]. GenAI offers a way to shift this model by creating more adaptive and personalized learning experiences. With its ability to respond in real time, support the creation of customized materials, and enable interactive, multimodal instruction, GenAI introduces a more dynamic and responsive approach to language education [28–30].

In a recent systematic review, Lee et al. [25] analyzed 49 empirical studies on classroom-based GenAI use in language education, highlighting its predominant focus on higher education and English language learners. Most studies emphasized learner perceptions–such as motivation, self-efficacy, and attitudes–while writing emerged as the most frequently addressed skill. GenAI tools were found to serve diverse roles, including feedback provider, tutor, and conversation partner. However, challenges such as content quality, information bias, and academic integrity remain pressing concerns. The review also identified critical gaps, including a lack of long-term and K–12 studies, limited behavioral analysis, and an over-reliance on self-reported data. These findings underscore the need for more robust, diverse, and pedagogically grounded research on GenAI's role in language education.

Tools like ChatGPT are among specific applications that have become well-known in English language teaching and learning. With the ability to create conversation models, offer real-time grammatical explanations, and even role-play in simulated conversations, Stockwell [31] investigate how ChatGPT serves as a linguistic resource and a teaching tool. Although they raise concerns about critical thinking, over-reliance, and the necessity of digital literacy, educators also report high learner engagement. A scoping literature review by Law [1] demonstrates that GenAI applications such as ChatGPT and GPT-based tools foster cognitive engagement, promote learner autonomy, and support constructive, output-oriented tasks. By using these tools, students are able to move beyond passive reception and generate meaningful written and spoken responses, such as engaging in purposeful dialogue and receiving contextualized feedback. Moreover, Ogunleye et al. [32] stress that tools like ChatGPT are most commonly used in

writing-intensive tasks such as to support brainstorming, content generation, and formative feedback. However, the authors also highlight the need for more structured pedagogical frameworks to ensure critical thinking and learning gains are sustained over time.

The study by Azamatova et al. [33] showed that combining AI tools with collaborative, task-based activities greatly increased students' motivation and achievement in foreign languages, demonstrating how well these tools fit with project-based learning approaches. When AI tools were added into the project cycle, learners not only did better on language tasks but also showed deeper engagement with the content.

A further major field of research examines how GenAI influences speaking anxiety and oral proficiency. Chen et al. [34] found that combining AI-driven speech recognition with dynamic assessment reduced cognitive load and linguistic anxiety while improving fluency and pronunciation. These results suggest that GenAI tools can create psychologically safe environments for language production, encouraging learner engagement. Building on this affective dimension, Wu and Liu [35] applied control-value theory to GenAI-supported L2 speaking practice. They found that enjoyment and curiosity significantly enhanced cognitive processing and performance, while GenAI competence predicted speaking success. Similarly, Wang et al. [36] used a latent profile analysis to examine engagement, self-efficacy, and anxiety in Chinese EFL learners using large language model-based tools, highlighting how different learner profiles interact with AI-supported instruction. Together, these studies highlight the role of psychological and motivational factors in shaping learners' engagement with GenAI tools in informal, autonomous contexts.

Other emerging studies also highlights the role of GenAI in supporting multilingual development and enhancing learner's sensitivity to cultural and contextual variation in language use. For example, Alharbi [37] provides a pedagogical overview of AI-based writing assistant tools such as machine translation, automated feedback, and GPT-based generation. This study highlights how GenAI tools help learners engage with diverse writing genres, refine their use of tone and register, and better address audience expectations in multilingual contexts. In a related study, Woo et al. [38] found that students who used AI-generated text as scaffolding showed improvements in text organization, vocabulary use, and syntactic complexity particularly in academic and professional writing tasks. These findings suggest that GenAI tools not only support language production but also promotes inclusive and differentiated learning by providing real-time, adaptive support for learners with diverse proficiency levels and cognitive profiles.

Nevertheless, important limitations persist in the application of GenAI in language education. Concerns about factual inaccuracies, hallucinations, and inconsistent feedback generated by LLMs raise questions about their reliability as instructional tools. For example, Han et al. [39] found that while ChatGPT was effective in supporting English as a foreign language learners during writing revision, it occasionally produced misleading or grammatically ambiguous suggestions. This suggest the underscoring the need for teacher oversight. Similarly, in a systematic review of 70 studies Lo et al. [40] noted that many GenAI tools lacked transparency in their output generation process, which posed challenges for learners in distinguishing between accurate and inaccurate information.

Overall, the emerging literature highlights a complex but optimistic outlook for the role of GenAI in language education. When applied with careful pedagogical planning,

GenAI technologies have the potential to strengthen language skills, reduce learner anxiety, and support independent and active learning. Nevertheless, important limitations remain: most studies focus on short-term interventions, single skills such as writing, or informal learning contexts, and they often rely heavily on self-reported outcomes. Furthermore, prior research has yet to fully explore how GenAI's unique affordances–such as natural language generation, adaptive feedback, and interactive dialogue–operate across diverse instructional settings, learner populations, and language skills. Addressing these gaps is essential for understanding the full potential of GenAI in language education, and motivates the present meta-analysis, which systematically evaluates its effects across formal and informal contexts

### 2.3 Moderators considered in this meta-analysis

When analyzing the impact of GenAI interventions on language learning outcomes, it is essential to consider a range of moderating variables that shape the tools' effectiveness across contexts. Moderators help explain heterogeneity in effect sizes and offer insights into the conditions under which GenAI can be most successfully implemented.

Recent meta-analytic work by Guan et al. [8] examined the role of GenAI in second language acquisition within self-directed, out-of-classroom settings. Their findings indicate that GenAI significantly supports gains in English proficiency and self-regulation, though its effect on learning motivation remains inconclusive, highlighting the need for motivation measures tailored to AI-enhanced environments. Building on this, the present study includes *Instructional Setting* as a moderator to compare the impact of GenAI in formal (classroom-based) versus informal (autonomous) learning contexts.

Lyu et al.[17] emphasized that both the presence of GenAI and the nature of comparison groups affect observed effect sizes. To ensure conceptual clarity, this meta-analysis only includes studies that explicitly involve GenAI tools and introduces *Design Type* as a moderator, distinguishing between experimental (control–treatment), quasi-experimental (non-randomized comparison), and within-subject (pre–post) designs, to assess how methodological choices influence reported outcomes.

As As Lee et al. [25] found in their systematic review, GenAI tools serve a variety of pedagogical roles in language education such as conversation partner, feedback generator, tutor, and content assistant. To reflect this functional diversity, we include *Intervention Type* as a moderator variable in our analysis. This allows us to evaluate whether different uses of GenAI produce varying effects on learning outcomes.

Furthermore, to capture the pedagogical and learner-specific variation as noted by Chen et al. [41], who identified three targeted aspects of AI-enabled assessment tools in language learning as language aspect moderators, we include *Language Skill* (e.g., writing, speaking, reading, listening) and *Educational Level* (e.g., primary, secondary, tertiary) as moderators. This approach aligns with Law [1] and Chen et al. [41], who also examined educational level as a moderator to determine how the effectiveness of GenAI interventions varies across different stages of formal education. As noted by Lee et al. [25], writing is by far the most commonly targeted skill in GenAI research, while other modalities remain underexplored. Their review also found a heavy concentration of studies in higher education, with limited coverage of K–12 settings. Including these moderators allows us to assess whether GenAI's impact varies depending on the specific linguistic competencies being targeted and the learner population involved.

In addition, we consider the *Economic Development* of the country in which each study was conducted, distinguishing between high-income and low and middle-income settings. This decision is motivated by findings from Mannuru et al. [42], who emphasize that GenAI adoption in developing countries is shaped by structural challenges such as digital access, infrastructure deficits, and teacher preparedness. At the same time, GenAI holds significant promise for expanding educational opportunities in under-resourced contexts. By including this moderator, we aim to explore whether the broader socio-economic context influences the effectiveness of GenAI interventions in language learning.

As highlighted in Chen et al. [41], variables such as target language, implementation environment, and intervention duration are commonly emphasized in studies examining the use of generative tools in K-12 language learning contexts. These variables are not only prominent in the literature but also facilitate consistent data analysis. Building on this, we include *Target Language*, *Learning Modality* (e.g., in-class, blended, or online), and *Intervention Duration* (e.g., short, medium, or long) as key moderators, respectively, to better understand how contextual and implementation-related factors influence the effectiveness of GenAI interventions.

We also included *Proficiency Level* as a moderator to account for differences in learners' linguistic backgrounds and competencies. For the purposes of our coding, we defined three levels based on educational stage and years of language study: beginner (up to a few years of study), intermediate (secondary school or early undergraduate learners), and advanced (graduate students or individuals with 10+ years of study). This classification aligns with the participant profiles described in previous research. For example, Liu [43] categorized participants based on their duration of English language learning and further verified their proficiency using national examination scores (Gaokao), which provided an objective benchmark for group-level comparisons.

Taken together, these moderators, including *Instructional Setting, Design Type, Intervention Type, Intervention Duration, Implementation Environment, Target Language, Language Skill, Educational Level, Proficiency Level, Learning Modality* allow a more nuanced analysis of GenAI's impact. However, prior research has rarely examined the combined influence of these factors in a systematic, quantitative way. By considering these contextual and methodological variables, the present meta-analysis addresses this gap and aims to clarify when, how, and for whom GenAI tools are most effective in second and foreign language education.

## 3 Research methods

Since the public release of ChatGPT in late 2022, there has been a rapid increase in empirical studies investigating the role of GenAI in language education. However, drawing overarching conclusions from this emerging body of work remains difficult due to variations in participant populations, language learning contexts, intervention types, and research designs. These differences contribute to inconsistent findings and limit generalizability. In this context, a meta-analytic approach is particularly valuable for aggregating results across studies, assessing overall trends, and identifying moderators that may explain variation in effectiveness [44, 45].

### 3.1 Study identification and selection

This study followed PRISMA 2020 guidelines [46] to ensure transparency and reproducibility in the selection process. A comprehensive search was conducted across four major academic databases—Scopus (75 articles), Web of Science (70), ACM Digital Library (6), and IEEE Xplore (4)—to identify peer-reviewed studies published from 2022 onwards. Reviews, conference papers, and non-English articles were excluded directly during the database searches by applying the relevant filters. Two additional articles were retrieved through backward snowballing. The full search strategy combined terms related to GenAI technologies and language education outcomes along with experimental design keywords (see full search string in Appendix A).

Studies were eligible for inclusion if they met the following criteria:

- Conducted in a second or foreign language (L2) learning context.
- Included a treatment group that used GenAI tools (e.g., GPT-based chatbots, AI tutors, LLM-based writing assistants).
- Employed a comparison group (e.g., traditional instruction, non-GenAI tools, or standard curriculum).
- Reported quantitative outcomes related to language learning (e.g., test scores, writing performance, speaking proficiency).
- Provided sufficient statistical information to compute effect sizes (or allowed for estimation from means, SDs, and sample sizes).

Figure 1 presents the study selection process, adapted from the PRISMA flow diagram. After removing duplicates and applying screening criteria, studies were excluded at various stages for several reasons, including the absence of a control group (n=5), lack of sufficient statistical information to compute effect sizes (n=12), or the use of outcome measures unrelated to language learning (n=3). In addition, 38 studies were removed due to methodological inconsistencies, lack of inclusion of specific GenAI tools (e.g., ChatGPT, Bard), absence of peer-reviewed publication, or because they focused on teacher

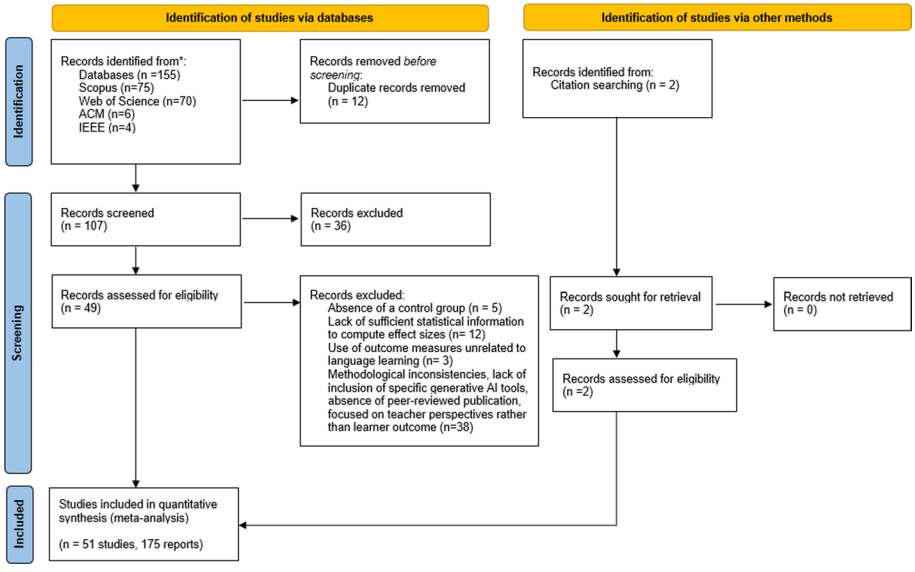



**Fig. 1** PRISMA flow diagram of the study selection process

perspectives rather than learner outcomes. In total, 51 eligible studies were identified, yielding a total of 175 effect size reports.

### 3.2 Effect size computation

To measure the effectiveness of GenAI interventions, Cohen's $d$ was used as the initial effect size metric, consistent with prior language education meta-analyses [47, 48]. When studies reported means and standard deviations for both treatment and control groups, Cohen's $d$ was computed as:

$$d = \frac{M_1 - M_2}{SD_{pooled}}, \quad \text{where} \quad SD_{pooled} = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}}$$

To correct for small sample bias, all reported effect sizes were adjusted using Hedges' $g$ [49], calculated as:

$$g = d \times J, \quad \text{where} \quad J = 1 - \frac{3}{4df - 1}, \quad df = n_1 + n_2 - 2$$

Effect sizes were synthesized using a random-effects model, which assumes that true effects vary across studies due to differences in participant populations, intervention types, and study contexts. This approach was chosen because substantial heterogeneity was expected among the included studies, consistent with prior meta-analyses in language education [8, 17]. Heterogeneity was further quantified using $Q$ statistics and $I^2$ values, and all reported effect sizes are expressed as Hedges' $g$ values.

All meta-analytic computations, including effect size calculations and adjustments, were conducted using Python (version 3.11.13) with the `statsmodels` package.

### 3.3 Data extraction and moderator coding

Data from each eligible study were systematically extracted and coded using a structured framework designed to support moderator analysis and effect size computation. Table 1 outlines the variables included in the coding process. Each study report was treated as a distinct entry and assigned identifiers, with outcomes categorized into either *Language Achievement* (e.g., reading, writing, speaking proficiency) or *Affective & Cognitive* outcomes (e.g., motivation, anxiety, engagement, self-regulation).

To maintain interpretive consistency, outcomes where a decrease is desirable–such as anxiety [50–54], boredom [55], hopelessness [56], or demotivation [57]–were reverse-coded so that all positive effect sizes represent favorable outcomes. Reverse coding was recorded as a separate binary moderator (*Reverse*) to allow for statistical checks of its influence in the meta-regression analyses. In two cases [58, 59], standard deviations were not reported; for these studies, they were imputed using the pooled standard deviation from other studies, following established practices for handling missing variance data in meta-analysis [60].

The selection of moderator variables was carefully guided by our literature review (see Section 2), which identified key theoretical and empirical factors influencing the effectiveness of GenAI in language education. As discussed in that section, prior meta-analyses in second language acquisition [8, 17, 18] and recent systematic reviews of GenAI in education [25, 42] have highlighted critical variables likely to shape intervention

**Table 1** Codebook for data extraction

| Variable | Description |
|---|---|
| Study (Report) | Identifier for each study outcome entry, including author(s), publication year, and outcome label if multiple outcomes are reported (e.g., *Xu and Liu (2025) – ChatGPT Critical Thinking*) |
| Bibkey | BibTeX citation key for linking entries to reference management software |
| Outcome Category | *Language Achievement:* performance-based outcomes (e.g., reading, writing, speaking, listening, grammar, overall proficiency); *Affective & Cognitive:* learner motivation, anxiety, engagement, satisfaction, self-regulated learning, etc. |
| Study Design | Type of study design: either *Experimental vs. Control* (between-subjects), *Pre-Post* (within-subjects), or *Quasi-experimental* (non-randomized comparison) |
| Specific Outcome | Specific measured outcome (e.g., speaking proficiency, motivation, anxiety) |
| Sample Sizes | Number of participants in each group: *Small* ($< 50$); *Medium* (50–200); *Large* ($> 200$) |
| Means and Standard Deviations | Reported means and standard deviations (or standard errors) for outcome measures |
| Language Skills | Specific aspect of language targeted: *Speaking*, *Reading*, *Writing*, *Listening*, or *Multiple skills* |
| Context | Type of language learning context: EFL, ESL, or other target languages |
| Target Language | The specific language being learned or taught in the study using the GenAI intervention (e.g., *English*, *Spanish*, *Korean*) |
| Educational Level | Participant educational level: *Primary*, *Secondary*, *Tertiary* (university), or *Adult* education |
| Instructional Setting | Whether the learning took place in a *Formal* (classroom-based, curriculum-driven) or *Informal* (self-directed, extracurricular) setting |
| Learning Modality | Mode of learning delivery: *In-Class*, *Online*, or *Blended* |
| Proficiency Level | *Beginner*: up to a few years of study; *Intermediate*: secondary school or early undergraduate learners; *Advanced*: graduate students or learners with 10+ years of English study |
| Intervention Type | Purpose for which AI tool was used: writing assistant, conversation partner, feedback generator, translation aid, speaking partner, etc. |
| Intervention Duration | *Short:* 1 session–5 weeks; *Medium:* 6–12 weeks; *Long:*$> 12$ weeks |
| Country | Country where the study was conducted |
| Economic Development | Whether the study was conducted in a *Developed* or *Developing* country (World Bank classification at publication time) |
| Continent | Continent where the study was conducted |
| Reverse | Binary variable for outcome direction: 0 = positive (e.g., proficiency, motivation); 1 = negative (e.g., anxiety, boredom) |

outcomes. These include *Instructional Setting* [8], *Intervention Type* (e.g., writing assistant, conversation partner, feedback generator) [25], *Language Skill* [41], *Educational Level* [25], and the *Economic Development* status of the country [42].

In some studies, there was ambiguity regarding the *Instructional Setting*. For example, GenAI tools were introduced in formal learning environments but students were encouraged to use them in informal, self-directed settings [61]. In such cases, the *Instructional Setting* was categorized as *informal* based on the primary context of tool usage. Conversely, when structured training or guided practice using GenAI tools was provided within a formal instructional setting, even if the main usage occurred outside the classroom, we classified the *Instructional Setting* as *formal* [62, 63]. Furthermore, following the methodological recommendations of Lyu et al. [17], we included *Design Type* (coded as *experimental vs. control*, *pre-post*, or *quasi-experimental*) to account for potential differences in the study design.

Each moderator variable was assigned a categorical code as specified in the codebook (Table 1). For example, *Educational Level* was coded as *primary*, *secondary*, *tertiary*, or *adult education*. Similarly, other categorical moderators, such as *Instructional Setting*, *Intervention Type*, and *Design Type*, were coded according to the definitions

operationalized in Table 1. To assess the robustness of moderator effects, we conducted sensitivity checks by excluding multi-level studies and re-running analyses.

Data extraction was carried out in a multi-stage, collaborative process involving all three authors. For each study, one author performed the initial coding, explicitly recorded in the codebook, and a second author independently verified the extracted data. For particularly complex moderators, a third author conducted an additional review to ensure accuracy. Any discrepancies were discussed and resolved through consensus among the reviewers. As a result, all coding decisions were fully agreed upon; formal inter-coder reliability statistics (e.g., Cohen's kappa) were not computed, since agreement was achieved through this collaborative process.

## 4 Results

This section presents the results of the meta-analysis in four parts. Section 4.1 reports the overall effects of GenAI on language learning outcomes. Section 4.2 provides a descriptive overview of the included studies. Section 4.3 addresses publication bias and small-study effects. Finally, Section 4.4 examines the role of key moderator variables in explaining variation in effect sizes.

### 4.1 Overall effects of GenAI on language education outcomes

The meta-analysis synthesized evidence from 51 primary studies, which together reported 175 independent effect sizes. These effect sizes were categorized into two primary outcome domains: *Language Achievement* and *Affective & Cognitive* outcomes. For *Language Achievement* outcomes, which included 71 effect sizes, the pooled effect size was large and statistically significant, with a Hedges' $g$ of 1.111 (95% CI [0.566, 1.656]). This indicates a substantial positive effect of GenAI tools on measurable language skills such as writing, speaking, reading, and vocabulary development.

Regarding *Affective & Cognitive* outcomes, encompassing 104 effect sizes, the pooled Hedges' $g$ was similarly large at 1.029 (95% CI [0.490, 1.567]). This demonstrates a meaningful positive influence of GenAI on learners' motivation, self-efficacy, metacognitive awareness, anxiety reduction, and related psychological constructs.

These findings suggest that while GenAI applications have robust positive effects on both language proficiency and affective-cognitive outcomes, the considerable heterogeneity observed underscores the need to interpret these effects in context. Moderator analyses were therefore essential to identify the study, learner, and intervention characteristics associated with stronger or weaker effects. Table 2 provides a concise summary of the pooled effect sizes and heterogeneity estimates for both outcome categories.

### 4.2 Characteristics of included studies

Table 3 provides an overview of the 51 unique studies analyzed to evaluate the effects of GenAI tools on language learning outcomes. About half of the studies (49.0%, e.g., [64–66]) employed an experimental vs. control group design, allowing for rigorous controlled comparisons. A substantial proportion (39.2%, e.g., [50, 67, 68]) used quasi-experimental

**Table 2** Summary of overall effects of GenAI interventions on language education outcomes

| Outcome category | k | Pooled $g$ | 95% CI | $I^2$ (%) |
|---|---|---|---|---|
| Language achievement | 71 | 1.111 | [0.566, 1.656] | 92.7 |
| Affective & cognitive | 104 | 1.029 | [0.490, 1.567] | 97.2 |

**Table 3** Summary of included studies

| Author(s), ref | Country | Learning modality | Design type | Interv. duration | Sample size | Language skill |
|---|---|---|---|---|---|---|
| Jaashan and Alashabi [64] | Saudi Arabia | In-class | Exp/Ctr | Short | Medium | Writing |
| Chen [50] | Taiwan | In-class | Quasi | Medium | Medium | Multiple |
| Alwasidi and Al-Khalifah [69] | Saudi Arabia | Blended | Pre/Post | Medium | Medium | Writing |
| Yilmaz and Aydin [67] | Türkiye | In-class | Quasi | Short | Small | Reading |
| Xu and Li [65] | China | In-class | Exp/Ctr | Medium | Medium | Multiple |
| Jamshed et al. [68] | India | In-class | Quasi | Long | Medium | Writing |
| Pan et al. [79] | China | Blended | Quasi | Medium | Medium | Reading |
| Ghafouri et al. [59] | Iran | In-class | Exp/Ctr | Medium | Small | Writing |
| Hwang et al. [66] | Taiwan | In-class | Exp/Ctr | Medium | Medium | Multiple |
| Alsofyani and Barzanji [80] | Saudi Arabia | In-class | Exp/Ctr | Medium | Medium | Writing |
| Darmawansah et al. [81] | Indonesia | Blended | Quasi | Medium | Medium | Speaking |
| Oktarina et al. [75] | Indonesia | In-class | Exp/Ctr | Long | Medium | Writing |
| Zheldibayeva [82] | Kazakhstan | Blended | Quasi | Medium | Medium | Listening |
| Kemelbekova et al. [83] | Kazakhstan | Blended | Exp/Ctr | Medium | Medium | Speaking |
| Yang et al. [71] | China | Blended | Quasi | Medium | Medium | Speaking |
| Kim [72] | South Korea | In-class | Exp/Ctr | Short | Small | Writing |
| Çelik et al. [70] | Türkiye | Blended | Pre/Post | Short | Medium | Reading |
| Chen [51] | Taiwan | Blended | Quasi | Medium | Medium | Multiple |
| Song and Song [84] | China | Blended | Exp/Ctr | Medium | Medium | Writing |
| Wu and Xu [55] | China | In-class | Pre/Post | Medium | Medium | Writing |
| Almineeai et al. [85] | Saudi Arabia | In-class | Pre/Post | Medium | Small | Speaking |
| Zahran [76] | Egypt | In-class | Quasi | Long | Medium | Multiple |
| Teng and Huan [77] | Japan | In-class | Exp/Ctr | Long | Medium | Writing |
| Zheng et al. [52] | China | In-class | Exp/Ctr | Short | Medium | Speaking |
| Ma and Chen [73] | China | Blended | Quasi | Short | Large | Multiple |
| Vicente-Yagüe-Jara et al. [74] | Spain | In-class | Quasi | Short | Medium | Writing |
| Liu and Wan [86] | China | In-class | Exp/Ctr | Medium | Medium | Multiple |
| Esfandiari and Allaf-Akbary [87] | Iran | Blended | Exp/Ctr | Medium | Medium | Writing |
| Boudouaia et al. [88] | Algeria | In-class | Exp/Ctr | Medium | Medium | Writing |
| Sayed et al. [89] | Ethiopia | In-class | Pre/Post | Short | Small | Speaking |
| Zheng [53] | China | In-class | Quasi | Short | Medium | Reading |
| Wang et al. [78] | China | Online | Exp/Ctr | Medium | Medium | Reading |
| Wiyaka et al. [54] | Indonesia | In-class | Exp/Ctr | Short | Medium | Reading |
| Wang and Akhter [56] | China | In-class | Exp/Ctr | Medium | Large | Multiple |
| Guo et al. [61] | China | In-class | Quasi | Short | Medium | Writing |
| Aldosari [90] | China | In-class | Quasi | Short | Medium | Speaking |
| Alsager [57] | Saudi Arabia | In-class | Exp/Ctr | Medium | Medium | Multiple |
| Esfandiari and Akbary [91] | Iran | In-class | Exp/Ctr | Short | Medium | Writing |
| Sapan and Uzun [92] | Türkiye | In-class | Quasi | Short | Medium | Writing |
| Urazbayeva et al. [63] | Kazakhstan | In-class | Pre/Post | Medium | Small | Multiple |
| Alnasib and Alharbi [62] | Saudi Arabia | In-class | Quasi | Long | Small | Writing |
| Meyer et al. [93] | Germany | In-class | Exp/Ctr | Short | Large | Writing |
| Ghafouri [58] | Iran | In-class | Exp/Ctr | Long | Small | Writing |
| Kucuk [94] | Iraq | In-class | Exp/Ctr | Medium | Medium | Multiple |
| Wu et al. [95] | Indonesia | Online | Quasi | Medium | Medium | Speaking |
| Liu et al. [96] | China | In-class | Exp/Ctr | Short | Medium | Writing |
| Kim and Su [97] | China | In-class | Quasi | Medium | Medium | Speaking |
| Kim et al. [98] | South Korea | In-class | Quasi | Long | Small | Writing |
| Zhou et al. [99] | China | In-class | Quasi | Medium | Medium | Speaking |
| Fathi et al. [100] | Iran | In-class | Exp/Ctr | Medium | Medium | Speaking |
| Rad et al. [101] | Iran | Blended | Exp/Ctr | Medium | Small | Writing |

*Exp/Ctr* Experimental vs. Control; *Quasi* Quasi-experimental

approaches, reflecting the prevalence of non-randomized comparisons in educational research. Finally, a smaller subset (11.8%, e.g., [55, 69, 70]) employed pre/post within-subject designs, which enable within-subject analyses but may be more vulnerable to confounding variables.

Geographically, research was concentrated predominantly in Asia, with over a third of studies (34.9%) originating from China (e.g., [52, 65, 71]), reflecting the region's rapid adoption and academic interest in GenAI applications. Other countries contributing to the literature include Saudi Arabia, Taiwan, Türkiye, Iran, and Indonesia, illustrating a diverse, though regionally clustered, research base.

The choice of GenAI tools varied, but ChatGPT was the most frequently studied, accounting for 55.6% of the included works (e.g., [59, 63, 72]). Other tools such as Replika, Gemini, ReadMate, and custom chatbots were also represented, suggesting a range of applications and developmental maturity among GenAI platforms in language education.

Studies examined a variety of learning contexts, including higher education settings, primary education, secondary education, and informal learning environments. The duration of the interventions ranged widely, from short-term studies (for example, a single session or a few weeks, [52, 73, 74]) to long-term studies that span entire academic terms (e.g., [75–77]), providing information on the immediate and sustained effects of GenAI use.

Several investigations focusing specifically on ChatGPT within higher education reported consistent positive impacts on diverse language learning outcomes such as writing proficiency, speaking fluency, and motivation [64, 67, 72]. Studies originating from China often combined ChatGPT with other digital tools like Duolingo or Typebot, leveraging multi-modal approaches for language and academic writing support [52, 65, 71, 78].

Overall, this body of research demonstrates a growing, geographically varied, and methodologically robust evidence base supporting the integration of GenAI tools in language education. The predominance of controlled experimental designs and the focus on popular, accessible tools like ChatGPT provide a solid foundation for meta-analytic synthesis.

Figure 2 illustrates the individual and pooled effect sizes for language achievement outcomes. The results indicate a significant overall benefit of GenAI interventions on language achievement. To enhance figure readability and avoid visual distortion, two studies with extreme outlier values were excluded from the figure but remained included in the meta-analytic model.

Figure 3 summarizes the effects on affective and cognitive outcomes, revealing a generally positive pattern, with most studies favoring GenAI interventions. Effect sizes varied across specific outcomes, but the overall trend pointed toward enhanced learner engagement, motivation, and cognitive strategy use. As with the previous figure, two extreme outliers were omitted from the plot to improve clarity, while retaining them in the meta-analysis.

### 4.3 Publication bias assessment

We visually inspected a funnel plot of effect sizes (Hedges *g*) against their standard errors to assess potential publication bias. Although some asymmetry was evident, the plot

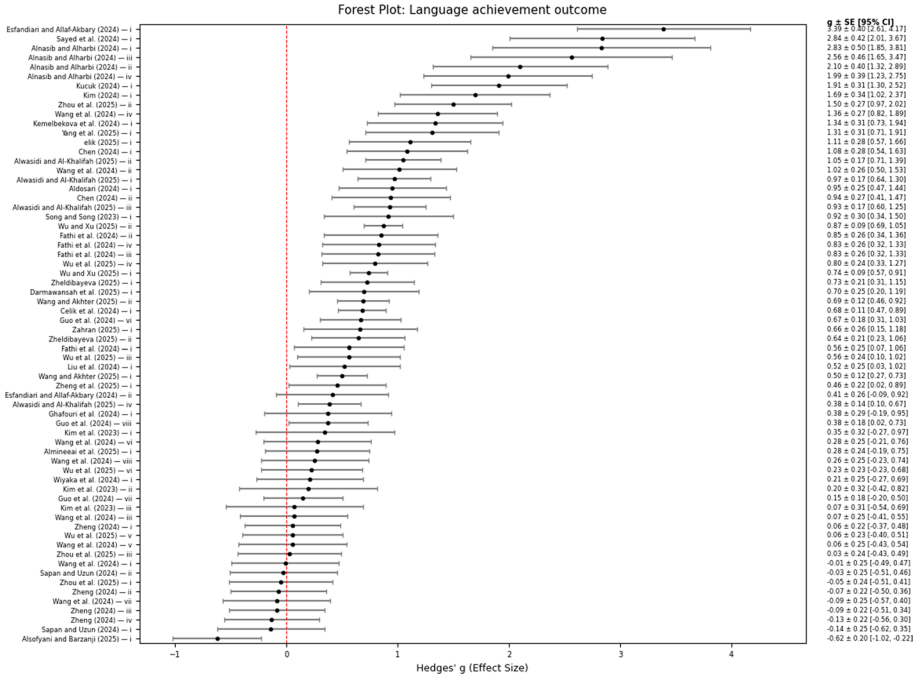



**Fig. 2** Forest plot showing the effect of GenAI on language achievement. Two studies with extreme effect size values were excluded from this figure to improve readability

displayed a reasonable spread of studies with varying effect sizes and standard errors, suggesting diversity in the reported outcomes. Overall, despite minor visual asymmetry, both quantitative and qualitative indicators suggest that publication bias is unlikely to have substantially affected the pooled effect size.

To quantitatively evaluate publication bias, Rosenthal's Fail-safe $N$ was computed [102], estimating the number of missing null-result studies required to reduce the overall effect size to non-significance at a two-tailed $\alpha$ level of 0.01:

$$N_{fs(0.01)} = \left(\frac{\sum Z}{2.33}\right)^2 - k,$$

where $Z$ represents the $Z$-value associated with each independent effect size, and $k$ is the total number of included studies. In this meta-analysis, $k = 175$, $\sum Z = 537.8$, and $N_{fs(0.01)} = 53,109$. The critical value for robustness is $5k + 10 = 885$, which is far exceeded, indicating minimal risk of substantial publication bias.

We assessed the methodological quality of included studies using the Cochrane Risk of Bias 2 (RoB 2) tool [103], which evaluates risk of bias across five domains: study design/confounding, missing data, outcome measurement, selective reporting, and other sources of bias. Details of the coding protocol are provided in Appendix C. Each study was rated as *low risk*, *some concerns*, or *high risk*, with overall risk categorized according to Cochrane guidance (low risk if all domains low, some concerns if $\geq 1$ domain raises concerns but none high, high risk if $\geq 1$ domain high). Detailed study-level assessments are provided in Table 4. Most studies were rated as low risk or some concerns, with relatively few studies categorized as high risk, indicating that the overall methodological quality was adequate but some caution is warranted.

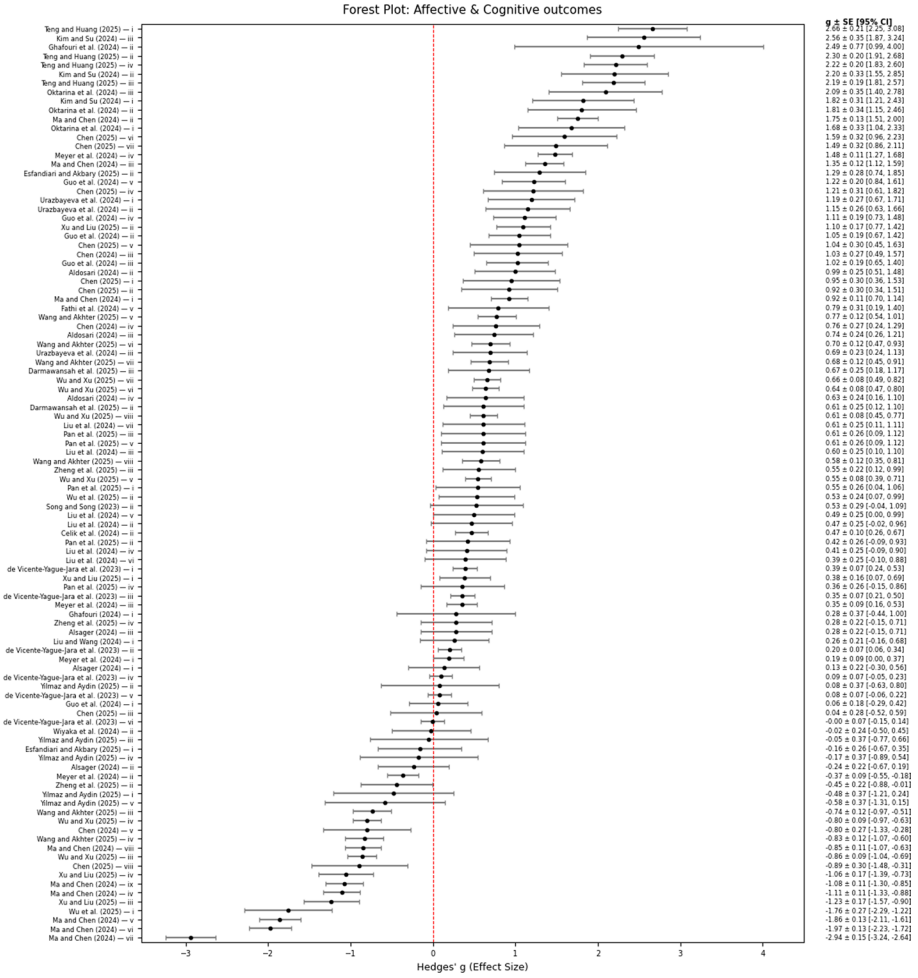



**Fig. 3** Forest plot showing the effect of GenAI on affective and cognitive outcomes. Two studies with extreme effect size values were excluded from this figure to improve readability

Sensitivity analyses were conducted to assess the robustness of the main findings. Excluding studies with small sample sizes ($N < 50$) reduced the dataset from 175 to 150 effect sizes. Excluding extreme effect sizes beyond $\pm 3$ SD reduced the dataset further to 147 effect sizes. In both cases, the overall meta-analytic conclusions remained consistent, indicating robustness to these potential sources of bias (see Table 5).

Egger's test indicated significant small-study effects (intercept = 3.06, $p < 0.001$), suggesting potential asymmetry in the funnel plot. However, trim-and-fill analysis imputed no missing studies, indicating minimal impact on the pooled effect size. The random-effects meta-analysis yielded a pooled Hedges $g = 0.81$ (95% CI: 0.61–1.02, $p < 0.001$). Heterogeneity across studies was substantial ($I^2 = 98.4\%$, $\tau^2 = 1.76$), indicating considerable variability in effect sizes.

In summary, the included studies generally exhibited adequate methodological quality, with most rated as low risk or some concerns, and sensitivity analyses indicate that the meta-analytic findings are robust to potential biases. Overall, the evidence appears reliable, though a small number of studies with higher risk or extreme effect sizes warrants cautious interpretation.

**Table 4** Risk-of-bias assessment for included studies (RoB 2.0 categories, two studies per row)

| Ref. | D | M | O | R | B | T | Ref. | D | M | O | R | B | T |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| [64] | L | L | SC | L | L | SC | [50] | SC | L | L | L | L | SC |
| [69] | H | L | L | L | L | H | [67] | SC | L | L | L | L | SC |
| [65] | L | L | L | L | L | L | [68] | SC | L | H | L | L | H |
| [79] | SC | L | L | L | L | SC | [59] | L | SC | L | L | L | SC |
| [66] | L | L | SC | L | L | SC | [80] | L | L | SC | L | L | SC |
| [81] | SC | L | L | L | L | SC | [75] | L | L | L | L | L | L |
| [82] | SC | L | L | L | L | SC | [83] | L | L | H | L | L | H |
| [71] | SC | L | L | L | L | SC | [72] | L | L | H | L | L | H |
| [70] | H | L | SC | L | L | H | [51] | SC | L | L | L | L | SC |
| [84] | L | L | L | L | L | L | [55] | H | L | L | L | L | H |
| [85] | H | L | SC | L | L | H | [76] | SC | L | L | L | L | SC |
| [77] | L | L | SC | L | L | SC | [52] | L | L | L | L | L | L |
| [73] | SC | L | L | L | L | SC | [74] | SC | L | L | L | L | SC |
| [86] | L | L | L | L | L | L | [87] | L | L | SC | L | L | SC |
| [88] | L | L | L | L | L | L | [89] | H | L | H | L | L | H |
| [53] | SC | L | L | L | L | SC | [78] | L | L | L | L | L | L |
| [54] | L | L | L | L | L | L | [56] | L | L | L | L | L | L |
| [61] | SC | L | L | L | L | SC | [90] | SC | L | L | L | L | SC |
| [57] | L | L | L | L | L | L | [91] | L | L | SC | L | L | SC |
| [92] | SC | L | SC | L | L | SC | [63] | H | L | H | L | L | H |
| [62] | SC | L | L | L | L | SC | [93] | L | L | SC | L | L | SC |
| [58] | L | SC | L | L | L | SC | [94] | L | L | SC | L | L | SC |
| [95] | SC | L | L | L | L | SC | [96] | L | L | L | L | L | L |
| [97] | SC | L | L | L | L | SC | [98] | SC | L | SC | L | L | SC |
| [99] | SC | L | L | L | L | SC | [100] | L | L | L | L | L | L |
| [101] | L | L | L | L | SC | SC | | | | | | | |

Column abbreviations: D = study design/confounding, M = missing data, O = outcome measurement, R = selective reporting, B = other biases, T = Total/overall risk-of-bias

Risk abbreviations: L = Low risk, SC = Some concerns, H = High risk

Overall coding: L if all domains are L; SC if $\geq$ 1 domain SC but none H; H if $\geq$ 1 domain H. Categories follow Cochrane RoB 2.0 guidance

### 4.4 Moderator analysis

A series of moderator analyzes revealed substantial and statistically significant heterogeneity in effect sizes across study characteristics (all $Q_{\text{between}}$ $p$-values < 0.001; see Table 6). The results indicate that the effect sizes varied systematically across several factors, including outcome category, study design, educational level, instructional setting, learning modality, proficiency level, type and duration of the intervention, sample size, targeted language ability, contextual factors and type of AI tool.

These patterns highlight meaningful differences between studies rather than a single homogeneous effect. Importantly, this heterogeneity suggests that reporting only a mean effect size may obscure critical variation and could lead to overgeneralization of the mean effect size. These findings underscore the importance of tailoring AI-supported language learning interventions to specific learner groups, instructional contexts, and AI tool types. Future research should explore additional sources of heterogeneity, possibly via meta-regression or more fine-grained subgroup analyses, to better understand which factors most strongly influence outcomes.

### 4.5 Reverse-coded outcome type as a moderator

The moderating effect of the outcome type's directionality was tested by categorizing outcomes as either positive (e.g., motivation, engagement) or negative (e.g., anxiety, inhibition). The test of moderation was statistically significant ($Q_{\text{between}} = 473.718$, $df = 1$, $p$ < 0.001), indicating a systematic difference in effect sizes between these two groups. Studies reporting positive outcomes ($k = 159$) showed a pooled effect size of $g = 0.432$ (SE = 0.013, 95% CI [0.406, 0.458]), while studies examining negative outcomes ($k = 16$) yielded a significantly higher pooled effect of $g = 0.601$ (SE = 0.041, 95% CI [0.522, 0.681]). These findings suggest that AI-based interventions may have a particularly strong influence on reducing negative psychological outcomes in language learning contexts.

## 5 Discussion

This meta-analysis synthesized the effects of GenAI tools on second and foreign language learning outcomes, drawing on 175 independent effect sizes across 51 studies. The findings indicate that GenAI interventions yield robust positive effects on both language achievement ($g$ = 1.111) and affective–cognitive outcomes ($g$ = 1.029), including reverse-coded outcomes such as learning anxiety, supporting the growing consensus that GenAI holds considerable promise for language education. However, the high degree of heterogeneity across studies also highlights the importance of context in determining the success of these tools. In what follows, we discuss these findings in relation to prior literature and key moderator effects, offering theoretical and pedagogical interpretations.

### 5.1 Comparison with prior meta-analyses

Our results align with and extend those reported in previous meta-analyses of AI in language learning. For instance, the large effect sizes observed here are consistent with the findings of Xiu-Yi [7], who reported a pooled SMD of 0.982 for general AI tools in L2 education, and Lyu et al. [17], who found medium-sized effects for chatbot-based interventions. However, our findings go further by isolating the specific contributions of GenAI tools, such as ChatGPT, Gemini, and AI-powered writing assistants. Notably,

**Table 5** Sensitivity analyses: effect sizes after excluding small studies and extreme values

| Sensitivity check | Number of effect sizes | Notes |
|---|---|---|
| Original dataset | 175 | All included studies |
| Exclude small studies (N < 50) | 150 | Small studies removed |
| Exclude extreme effect sizes (±3 SD) | 147 | Extreme effect sizes removed |

the effect sizes in our study exceed those reported by Guan et al.[8] for GenAI use in informal language learning ($g$ = 0.447), likely due to our broader inclusion of outcome categories and more diverse intervention types.

### 5.2 Intervention duration and instructional context

Among all moderators tested, intervention duration emerged as one of the strongest predictors of effectiveness. Long-term interventions produced markedly larger effect sizes ($g$ = 1.668) than medium- or short-term ones. This dose–response relationship supports the findings of Xiu-Yi [7], who also observed stronger effects for medium-duration studies and argued for the pedagogical importance of sustained AI integration. In the context of GenAI, this may be particularly relevant, as extended exposure allows learners to develop familiarity with tools, benefit from iterative feedback, and engage in deeper forms of language practice and reflection.

Similarly, we found that GenAI interventions in informal settings (e.g., autonomous or extracurricular use) were significantly more effective ($g$ = 0.741) than those implemented in formal, classroom-based contexts ($g$ = 0.436). This replicates the pattern observed by Guan et al. [8] and supports the argument that GenAI may be especially well-suited for self-directed learning. Informal settings likely afford greater learner agency, flexibility, and reduced performance pressure–factors that enhance motivation and allow learners to experiment with language use more freely [28, 34].

### 5.3 Tool type and functional roles

The type of GenAI intervention was also a significant moderator, with feedback-generating tools ($g$ = 1.708) and productivity assistants ($g$ = 1.454) yielding the highest effects. These findings echo qualitative insights from Law [1] and Ogunleye et al.[32], who emphasized the value of GenAI tools in facilitating output-focused tasks and supporting learner autonomy. The substantial gains associated with feedback tools are especially noteworthy, as they reinforce the idea that timely, individualized responses–hallmarks of GenAI–can significantly enhance language learning outcomes, particularly in writing-intensive contexts [33].

### 5.4 Learner characteristics and skill focus

Effects also varied by learner proficiency and educational level. Advanced learners benefited more from GenAI use than beginners or intermediates, possibly because they are better equipped to leverage AI-generated input for self-regulated learning, as noted by Alharbi [37]. Moreover, adults and tertiary-level learners showed larger gains than K–12 learners, consistent with Lee et al. [25], who found a research bias toward higher education settings and reported that university students tend to have more positive perceptions of GenAI tools.

Among the language skills targeted, listening ($g$ = 1.050), speaking ($g$ = 0.686), and writing ($g$ = 0.545) showed the highest gains. These results support the notion that

**Table 6** Moderator analyses of the pooled effect sizes of GenAI interventions in language learning

| Moderator | k | Pooled g | SE | 95% CI | | Z-value | p-value |
|---|---|---|---|---|---|---|---|
| | | | | LL | UL | | |
| *Outcome category* $Q_{between}$ = 6652.451 df = 1 p < 0.001 | | | | | | | |
| Affective & cognitive | 104 | 0.382 | 0.015 | 0.353 | 0.410 | 26.147 | 0.000e + 00 |
| Language achievement | 71 | 0.651 | 0.026 | 0.601 | 0.701 | 25.529 | 0.000e + 00 |
| *Design type* $Q_{between}$ = 931.906 df = 1 p < 0.001 | | | | | | | |
| Exp/control | 73 | 0.596 | 0.023 | 0.552 | 0.640 | 26.315 | 0.000e + 00 |
| Pre/Post(within-subject) | 19 | 0.711 | 0.026 | 0.661 | 0.762 | 27.594 | 0.000e + 00 |
| Quasi | 83 | 0.202 | 0.019 | 0.164 | 0.239 | 10.608 | 0.000e + 00 |
| *Educational level* $Q_{between}$ = 320.925 df = 3, p < 0.001 | | | | | | | |
| Adult | 6 | 0.839 | 0.114 | 0.616 | 1.063 | 7.361 | 1.825e-13 |
| Primary | 7 | 0.499 | 0.095 | 0.312 | 0.685 | 5.234 | 1.657e-07 |
| Secondary | 14 | 0.303 | 0.041 | 0.223 | 0.383 | 7.413 | 1.237e-13 |
| Tertiary | 148 | 0.458 | 0.014 | 0.431 | 0.484 | 33.764 | 0.000e + 00 |
| *Learning modality* $Q_{between}$ = 8415.257 df = 2 p < 0.001 | | | | | | | |
| Blended | 42 | 0.149 | 0.028 | 0.094 | 0.204 | 5.300 | 1.158e-07 |
| In-class | 119 | 0.529 | 0.015 | 0.500 | 0.557 | 36.335 | 0.000e + 00 |
| Online | 14 | 0.454 | 0.066 | 0.325 | 0.583 | 6.898 | 5.286e-12 |
| *Instructional setting* $Q_{between}$ = 359.306 df = 1 p < 0.001 | | | | | | | |
| Formal | 165 | 0.436 | 0.013 | 0.411 | 0.462 | 33.721 | 0.000e + 00 |
| Informal | 10 | 0.741 | 0.063 | 0.617 | 0.866 | 11.683 | 0.000e + 00 |
| *Proficiency level* $Q_{between}$ = 4176.489 df = 3 p < 0.001 | | | | | | | |
| Advanced | 19 | 0.715 | 0.038 | 0.642 | 0.789 | 18.986 | 0.000e + 00 |
| Beginner | 22 | 0.466 | 0.047 | 0.374 | 0.558 | 9.929 | 0.000e + 00 |
| Intermediate | 87 | 0.365 | 0.016 | 0.334 | 0.396 | 23.045 | 0.000e + 00 |
| Not explicitly reported | 47 | 0.575 | 0.030 | 0.515 | 0.634 | 18.882 | 0.000e + 00 |
| *Intervention type* $Q_{between}$ = 4879.169 df = 6 p < 0.001 | | | | | | | |
| Conversation partner | 2 | 0.577 | 0.208 | 0.169 | 0.985 | 2.771 | 5.589e-03 |
| Feedback generator | 9 | 1.708 | 0.078 | 1.555 | 1.861 | 21.889 | 0.000e + 00 |
| Listning partner | 1 | 0.726 | 0.214 | 0.306 | 1.146 | 3.390 | 6.981e-04 |
| Multiple | 109 | 0.384 | 0.014 | 0.356 | 0.412 | 27.274 | 0.000e + 00 |
| Speaking partner | 27 | 0.549 | 0.049 | 0.453 | 0.645 | 11.215 | 0.000e + 00 |
| Translation aid | 2 | 0.567 | 0.075 | 0.421 | 0.714 | 7.613 | 2.687e-14 |
| Writing assistant | 25 | 0.569 | 0.052 | 0.468 | 0.671 | 11.012 | 0.000e + 00 |
| *Intervention duration* $Q_{between}$ = 32,908.470 df = 2 p < 0.001 | | | | | | | |
| Long | 19 | 1.668 | 0.066 | 1.540 | 1.797 | 25.465 | 0.000e + 00 |
| Medium | 94 | 0.649 | 0.018 | 0.613 | 0.685 | 35.071 | 0.000e + 00 |
| Short | 62 | 0.165 | 0.018 | 0.130 | 0.200 | 9.140 | 0.000e + 00 |
| *AI tool group* $Q_{between}$ = 99.985 df = 2 p < 0.001 | | | | | | | |
| LLM tools | 153 | 0.448 | 0.013 | 0.422 | 0.473 | 34.498 | 0.000e + 00 |
| Specialized tools | 18 | 0.387 | 0.061 | 0.267 | 0.507 | 6.327 | 2.502e-10 |
| Productivity assistants | 4 | 1.454 | 0.223 | 1.018 | 1.890 | 6.531 | 6.546e-11 |
| *Sample size* $Q_{between}$ = 8631.449 df = 2 p < 0.001 | | | | | | | |
| Large | 21 | 0.148 | 0.025 | 0.099 | 0.197 | 5.936 | 2.917e-09 |
| Medium | 129 | 0.542 | 0.015 | 0.513 | 0.572 | 35.996 | 0.000e+00 |
| Small | 25 | 0.781 | 0.069 | 0.645 | 0.916 | 11.288 | 0.000e+00 |
| *Language skill* $Q_{between}$ = 8135.827 df = 4 p < 0.001 | | | | | | | |
| Listening | 2 | 1.050 | 0.194 | 0.669 | 1.431 | 5.403 | 6.544e-08 |
| Multiple skills | 47 | 0.198 | 0.024 | 0.150 | 0.246 | 8.084 | 6.661e-16 |
| Reading | 27 | 0.360 | 0.043 | 0.276 | 0.444 | 8.377 | 0.000e+00 |
| Speaking | 31 | 0.686 | 0.046 | 0.595 | 0.776 | 14.822 | 0.000e+00 |
| Writing | 68 | 0.545 | 0.017 | 0.512 | 0.578 | 32.330 | 0.000e+00 |
| *Target language* $Q_{between}$ = 3330.143 df = 2 p < 0.001 | | | | | | | |
| English | 166 | 0.506 | 0.014 | 0.479 | 0.534 | 36.060 | 0.000e+00 |

**Table 6** (continued)

| Moderator | k | Pooled g | SE | 95% CI | | Z-value | p-value |
|---|---|---|---|---|---|---|---|
| | | | | LL | UL | | |
| Korean | 3 | 0.704 | 0.191 | 0.331 | 1.078 | 3.696 | 2.192e-04 |
| Spanish | 6 | 0.181 | 0.030 | 0.123 | 0.239 | 6.077 | 1.227e-09 |
| *Economic development* $Q_{between}$ = 176.151 df = 1 p < 0.001 | | | | | | | |
| Developed | 33 | 0.416 | 0.023 | 0.371 | 0.462 | 17.894 | 0.000e+00 |
| Developing | 142 | 0.462 | 0.015 | 0.432 | 0.492 | 30.559 | 0.000e+00 |
| *Continent* $Q_{between}$ = 5371.432 df = 2 p < 0.001 | | | | | | | |
| Africa | 3 | 1.784 | 0.216 | 1.359 | 2.208 | 8.242 | 2.220e-16 |
| Asia | 153 | 0.524 | 0.015 | 0.495 | 0.554 | 34.776 | 0.000e+00 |
| Europe | 19 | 0.247 | 0.024 | 0.201 | 0.294 | 10.502 | 0.000e+00 |
| *Reverse* $Q_{between}$ = 473.718 df = 1 p < 0.001 | | | | | | | |
| 0 | 159 | 0.432 | 0.013 | 0.406 | 0.458 | 32.357 | 0.000e+00 |
| 1 | 16 | 0.601 | 0.041 | 0.522 | 0.681 | 14.838 | 0.000e+00 |

GenAI is particularly effective for productive and interactive skills, where its generative and conversational capabilities can be fully leveraged [31, 38]. The strong results for speaking are also consistent with studies like Chen et al.[34], which found that GenAI-based tools can reduce speaking anxiety and enhance fluency through low-stakes, supportive interaction.

### 5.5 Addressing heterogeneity and contextual factors

Despite the overall positive findings, our results revealed substantial heterogeneity, pointing to the critical role of contextual variables. For example, studies conducted in developing countries and in African contexts reported some of the largest effect sizes. These findings resonate with Mannuru et al. [42], who argue that GenAI may offer especially transformative benefits in resource-constrained settings, where access to high-quality educational support is limited. However, they also underscore the need for infrastructure, teacher training, and digital equity to support sustainable adoption.

The presence of small-study effects—-where smaller-sample studies tended to report larger effects—-calls for caution in interpreting effect magnitudes. While publication bias analyses indicated robust findings (see Section 4.3 for details), indicated robust findings, the inverse relationship between sample size and effect size suggests potential small-study effects [104]. A few studies reported unusually large effect sizes that may have inflated the overall estimates. For example, Rad et al. [101] reported two exceptionally high effect sizes for writing performance and writing engagement. Although these were retained to preserve data integrity, their influence underscores the importance of interpreting pooled estimates with caution and contextual nuance.

### 5.6 Theoretical and pedagogical implications

Theoretically, these findings are consistent with sociocognitive and constructivist perspectives on language learning, which emphasize interaction, feedback, and learner agency [14, 15]. In particular, GenAI tools–especially when used over longer periods and in learner-directed contexts–appear to enhance opportunities for authentic communication and scaffolded output, core principles of task-based and communicative language teaching frameworks [28, 33]. Moderator analyses further indicate that interventions providing feedback or conversational practice yielded larger effects, supporting the notion that guided interaction is critical for language development. These results

also align with recent work by Wang et al. [10], who developed and validated a GenAI literacy scale for EFL teachers, highlighting the importance of teacher competence in scaffolding effective AI-supported learning interactions.

Pedagogically, the results suggest that GenAI is not a "one size fits all" solution but must be aligned with learners' proficiency levels, educational goals, and institutional contexts. Long-term integration, use in informal or hybrid settings, and strategic deployment of GenAI as feedback or writing support can maximize its impact. Moreover, interventions using AI-driven chatbots have been shown to enhance student engagement in EFL contexts [105], reinforcing the potential of conversational AI to support active, learner-centered pedagogy. Teachers remain central in guiding students' interaction with GenAI tools, ensuring content reliability, and fostering critical engagement with AI-generated input.

### 5.7 Limitations and future research

Several limitations of this meta-analysis should be acknowledged. First, the high heterogeneity across studies complicates efforts to generalize findings. Part of this heterogeneity likely arises from our aggregation of diverse outcome measures under the two broad categories of *Language Achievement* and *Affective & Cognitive* outcomes. While this approach provides a comprehensive overview, pooling such varied measures naturally contributes to the very high heterogeneity observed in the analyses (as indicated by the $Q$ and $I^2$ values).

Second, some moderator subgroups (e.g., reverse-coded outcomes, less common intervention types) included relatively few studies, which may limit the robustness of these specific findings. Although publication bias analyses indicated minimal impact on the overall results, caution is warranted when interpreting results for smaller subsets of studies.

In addition to the limitations of the meta-analysis itself, it is important to recognize constraints of the primary studies on which it relies. Most included studies focused on English as the target language and on adult or university learners, limiting the applicability of results to younger learners or less commonly taught languages. While the review covered a wide array of outcome types, most studies relied on self-reported or short-term measures. This pattern is not unexpected, as all included studies were conducted between 2023 and 2025, shortly after the introduction of ChatGPT and other widely accessible GenAI tools, leaving limited opportunity for longitudinal designs to emerge.

Another limitation concerns the geographical distribution of the included studies summarized in Table 3. The majority were conducted in Asian contexts, with little representation from Western regions and none from the United States or the United Kingdom. This distribution reflects the current empirical literature available that met the predefined inclusion criteria (e.g., GenAI interventions with pre- and post-test outcome measures), rather than selective exclusion of particular regions. Notably, a recent systematic review of GenAI use in language classrooms by Lee et al. [25] reported a highly similar geographical pattern, with the majority of studies conducted in Asia and only a small number originating from the United States or Europe. Prior reviews attribute this concentration to regional research priorities, institutional infrastructure, and early large-scale adoption of GenAI tools in higher education, particularly in East Asia.

Nevertheless, this geographical imbalance may limit the generalizability of the present findings across diverse educational systems and sociocultural contexts.

Future research should therefore prioritize longer-term investigations, objective assessments, and more diverse learner populations across a wider range of geographical contexts. As GenAI-based language learning interventions become more widely implemented and systematically evaluated across different educational systems, future meta-analyses may revisit these questions using a broader and more geographically balanced evidence base. Moreover, as noted by Han et al. [39] and Lo et al. [40], the reliability of GenAI-generated content remains an issue. Continued emphasis on digital literacy and teacher oversight will be critical as GenAI tools become more embedded in language education.

## 6 Conclusion

This meta-analysis synthesized findings from 51 studies reporting 175 independent effect sizes to evaluate the impact of GenAI tools on second and foreign language learning. While individual studies and conceptual reviews have highlighted GenAI's promise [28, 31], this is the first quantitative synthesis to assess its effectiveness across instructional contexts and learner populations.

Overall, GenAI tools exert a large, positive effect on language learning outcomes. Compared to traditional instructional approaches or control conditions, GenAI-enhanced interventions significantly improved both language achievement ($g$ = 1.111) and affective–cognitive outcomes such as motivation, engagement, and self-efficacy ($g$ = 1.029). These effects exceed those reported in previous meta-analyses on general AI use in language education [7, 17], underscoring the transformative potential of GenAI when effectively implemented.

Moderator analyses further revealed that GenAI's impact varies by learner, context, and intervention characteristics. Stronger effects were observed in informal settings, longer interventions, and studies involving adult or higher-proficiency learners. Feedback-generating tools and productivity assistants were particularly effective for productive skills such as speaking and writing, supporting interactionist and sociocultural perspectives that emphasize scaffolded input, feedback, and learner autonomy [33, 34].

Notably, studies targeting less commonly taught languages (e.g., Korean) and those conducted in developing regions reported especially high effects, suggesting that GenAI may help bridge access and equity gaps in global language education [42]. However, high heterogeneity and the presence of small-study effects indicate that these benefits are not uniformly distributed, with a few studies (e.g., [101]) reporting unusually large effects that may have modestly inflated pooled estimates.

In conclusion, these findings support a sociocognitive view of GenAI as an interactive scaffold that enhances learner agency and output. Practically, they underscore that GenAI should not be treated as a "plug-and-play" solution but tailored to specific educational contexts, learner needs, and skill targets. Future research should diversify languages, expand geographic coverage, and explore long-term and transfer effects to ensure equitable and pedagogically grounded integration of GenAI in language education.

## Appendix A: Search string used for literature retrieval

The full Boolean search string used across databases (Scopus, Web of Science, ACM DL, IEEE Xplore) was:

> *("generative artificial intelligence" OR GenAI OR"ChatGPT" OR "Claude" OR "Gemini" OR chatbot* OR "conversational agent*" OR GPT* OR "large language model*" OR LLM OR "transformer-based model*" OR "AI writing assistant*" OR "AI-powered chatbot*" OR "AI language tutor*" OR "AI-based language model*" OR "text-generating AI") AND ("foreign language education" OR "second* language education" OR "L2 education" OR "foreign language learning" OR "second* language learning" OR "language teaching" OR "language instruction" OR "English as a Foreign Language" OR "English as a Second Language") AND (pretest OR posttest OR "control group" OR "comparison group" OR "treatment group" OR "experimental")*

## Appendix B: Funnel plot

For completeness, Fig. 4 presents the funnel plot of all included effect sizes. As discussed in Sect. 4.3, Egger's test indicated significant small-study effects, but trim-and-fill suggested minimal impact on the pooled estimate.

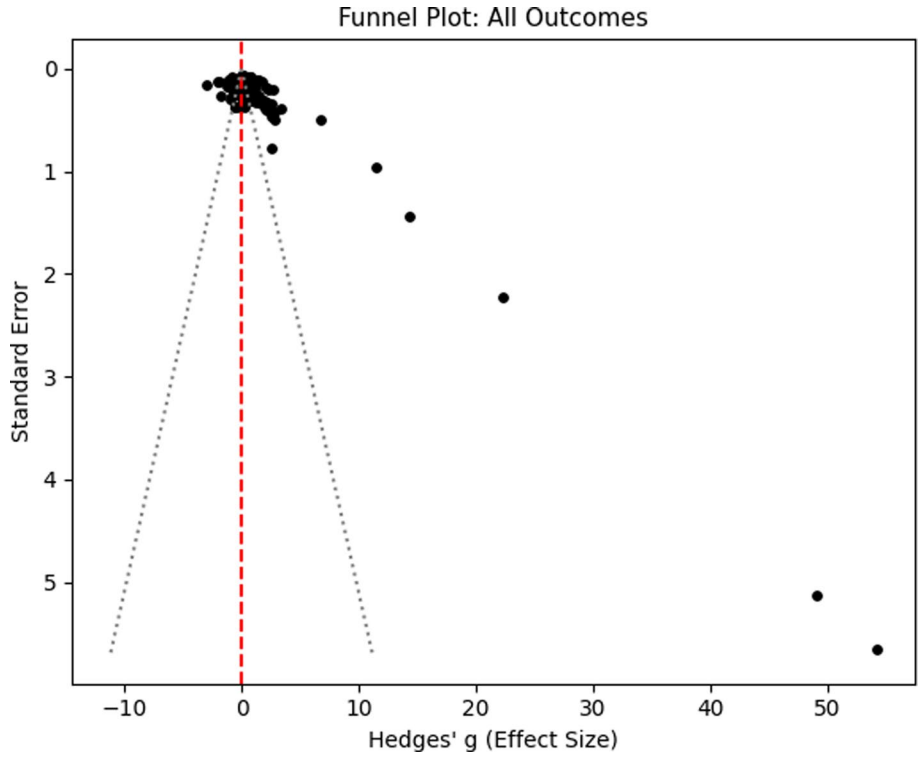



**Fig. 4** Funnel plot for all included effect sizes

## Appendix C: Coding protocol for risk-of-bias assessment

**Table 7** Coding protocol for risk-of-bias assessment (RoB 2.0)

| RoB 2.0 Domain | Criteria for rating | Coding |
|---|---|---|
| Study design/ confounding (D) | L: True experimental/randomized controlled study<br>SC: Quasi-experimental study<br>H: Single-group pre–post study | L/ SC/H |
| Missing data/ attrition (M) | L: Complete data reported<br>SC: Minor missing data (e.g., missing standard deviation)<br>H: Major missing data likely to bias results | L/ SC/H |
| Outcome measurement (O) | L: Validated instruments (TOEFL iBT, IELTS rubrics, established psychometric scales)<br>SC: Adapted validated instruments (modified rubrics, checklists based on curriculum, partial UTAUT scales)<br>H: Unvalidated or self-developed instruments | L/ SC/H |
| Selective reporting (R) | L: All pre-specified outcomes reported<br>SC: Minor outcomes missing<br>H: Major outcomes missing or selectively reported | L/ SC/H |
| Other biases (B) | L: No obvious other sources of bias<br>SC: Minor concerns (small imbalances, small sample)<br>H: Major concerns (extreme effect sizes, conflicts of interest) | L/ SC/H |
| Overall risk-of-bias (T) | L: All domains L<br>SC: $\geq$1 domain SC but none H<br>H: $\geq$1 domain H | L/ SC/H |

Risk abbreviations: L = Low risk, SC = Some concerns, H = High risk

#### Author contributions

M.S., S.G., P.K. wrote the main manuscript text, M.S. conceptualized the article and conducted the analysis, S.G., P.K., M.S. collected the data, All authors reviewed the manuscript.

#### Funding

Open Access funding provided by University of Jyväskylä (JYU). This work was supported by the Research Council of Finland (project no. 356314).

#### Data availability

No datasets were generated or analysed during the current study.

### Declarations

#### Ethics approval and consent to participate

The authors of this manuscript confirm that ethical approval and informed consent not required as this is a review paper.

#### Consent for publication

The authors of this manuscript confirm that consent for publication is not required as no individual person's data is included in this review.

#### Competing interests

The authors declare no competing interests.

Received: 21 July 2025 / Accepted: 16 February 2026

Published online: 24 February 2026

#### References

1. Law L. Application of generative artificial intelligence (GenAI) in language teaching and learning: a scoping literature review. Computers and Education Open. 2024;6:100174.
2. Pack A, Maloney J. Using generative artificial intelligence for language education research: insights from using OpenAI's ChatGPT. TESOL Q. 2023;57(4):1571–82.
3. Derakhshan A, Ghiasvand F. Is ChatGPT an evil or an angel for second language education and research? A phenomenographic study of research-active EFL teachers' perceptions. Int J Appl Linguist. 2024;34(4):1246–64.
4. Elyoseph Z, Hadar-Shoval D, Asraf K, Lvovsky M. ChatGPT outperforms humans in emotional awareness evaluations. Front Psychol. 2023;14:1199058.

5. Wang J, Fan W. The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis. Humanit Soc Sci Commun. 2025;12(1):1–21.
6. Chapelle CA. Generative ai as game changer: implications for language education. System. 2025;132:103672.
7. Xiu-Yi W. AI in L2 learning: a meta-analysis of contextual, instructional, and social-emotional moderators. System. 2024;126:103498.
8. Guan L, Li S, Gu MM. AI in informal digital English learning: a meta-analysis of its effectiveness on proficiency, motivation, and self-regulation. Computers and Education Artificial Intelligence. 2024;7:100323.
9. Nguyen Thi Thu, H. EFL teachers' perspectives toward the use of ChatGPT in writing classes: a case study at Van Lang University. Nguyen, TTH. EFL teachers' perspectives toward the use of ChatGPT in writing classes: a case study at Van Lang University. Int J Lang Instr. 2023;2(3):1–47.
10. Wang Y, Derakhshan A, Ghiasvand F. EFL teachers' generative artificial intelligence (GenAI) literacy: a scale development and validation study. System. 2025;133:103791.
11. Wu H, Dong Z. What motivates second language majors to use generative AI for informal learning? Insights from the theory of planned behavior. IEEE Access. 2025;13:34877–86.
12. Fryer L, Carpenter R. Bots as language learning tools. Lang Learn Technol. 2006;10(3):8–14.
13. Huang W, Hew KF, Fryer LK. Chatbots for language learning-are they really useful? A systematic review of chatbot-supported language learning. J Comput Assist Learn. 2022;38(1):237–57.
14. Vygotsky LS, Cole M. Mind in society: development of higher psychological processes. Cambridge, MA: Harvard University Press; 1978.
15. Lantolf JP, Thorne SL, Poehner ME. Sociocultural theory and second language development. In: Theories in second language acquisition. 3rd edn. 2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN and 52 Vanderbilt Avenue, New York, NY 10017: Routledge; 2020. p. 221–40.
16. Long MH. The role of the linguistic environment in second language acquisition. In: Ritchie WC, Bhatia TK, editors. Handbook of second language acquisition. New York: Academic Press; 1996. p. 413–68.
17. Lyu B, Lai C, Guo J. Effectiveness of chatbots in improving language learning: a meta-analysis of comparative studies. Int J Appl Linguist. 2025;35(2):834–51.
18. Lee S-M. The effectiveness of machine translation in foreign language education: a systematic review and meta-analysis. Comput Assist Lang Learn. 2023;36(1–2):103–25.
19. Xu Z, Wijekumar K, Ramirez G, Hu X, Irey R. The effectiveness of intelligent tutoring systems on K-12 students' reading comprehension: a meta-analysis. Br J Educ Technol. 2019;50(6):3119–37.
20. Underwood J. Simulated conversation as a CAI strategy. Foreign Lang Ann. 1982;15(3):209–12.
21. Heift T. Multiple learner errors and meaningful feedback. CALICO J. 2003;20(3):533–48.
22. Warschauer M, Grimes D. Automated writing assessment in the classroom. Pedagogies Int J. 2008;3(1):22–36.
23. Nyns RR. Is intelligent computer-assisted language learning possible? System. 1989;17(1):35–47.
24. Huang F, Wang Y, Zhang H. Modelling generative ai acceptance, perceived teachers' enthusiasm and self-efficacy to English as a foreign language learners' well-being in the digital era. Eur J Educ. 2024;59(4):12770.
25. Lee S, Choe H, Zou D, Jeon J. Generative AI (GenAI) in the language classroom: a systematic review. Interact Learn Environ. 2025. https://doi.org/10.1080/10494820.2025.2498537.
26. Razafinirina MA, Dimbisoa WG, Mahatody T. Pedagogical alignment of large language models (LLM) for personalized learning: a survey, trends and challenges. J Intell Learn Syst Appl. 2024;16(4):448–80.
27. Zhou G, Niu X. Approaches to language teaching and learning. J Lang Teach Res. 2015;6(4):798.
28. Creely E. Exploring the role of generative AI in enhancing language learning: opportunities and challenges. Int J Changes Educ. 2024;1(3):158–67.
29. Baidoo-Anu D, Ansah LO. Education in the era of generative artificial intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning. J AI. 2023;7(1):52–62.
30. Vandewaetere M, Clarebout G. Advanced technologies for personalized learning, instruction, and performance. In: Spector JM, Merrill MD, Elen J, Bishop MJ, editors. Handbook of research on educational communications and technology. New York, NY: Springer; 2013. p. 425–37.
31. Stockwell G. ChatGPT in language teaching and learning: exploring the road we're travelling. Technol Lang Teach Learn. 2024;6(1):2273–2273.
32. Ogunleye B, Zakariyyah KI, Ajao O, Olayinka O, Sharma H. A systematic review of generative AI for teaching and learning practice. Educ Sci. 2024;14(6):636.
33. Azamatova A, Bekeyeva N, Zhaxylikova K, Sarbassova A, Ilyassova N. The effect of using artificial intelligence and digital learning tools based on project-based learning approach in foreign language teaching on students' success and motivation. Int J Educ Math Sci Technol. 2023;11(6):1458–75.
34. Chen C-H, Koong C-S, Liao C. Influences of integrating dynamic assessment into a speech recognition learning design to support students' English speaking skills, learning anxiety and cognitive load. Educ Technol Soc. 2022;25(1):1–14.
35. Wu H, Liu W. Exploring mechanisms of effective informal GenAI-supported second language speaking practice: a cognitive-motivational model of achievement emotions. Discov Comput. 2025;28(1):119.
36. Wang Q, Gao Y, Wang X. Exploring engagement, self-efficacy, and anxiety in large language model EFL learning: a latent profile analysis of chinese university students. Int J Hum Comput Interact. 2025;41(12):7815–24.
37. Alharbi W. AI in the foreign language classroom: a pedagogical overview of automated writing assistance tools. Educ Res Int. 2023;2023(1):4253331.
38. Woo DJ, Susanto H, Yeung CH, Guo K, Fung AKY. Exploring AI-generated text in student writing: how does AI help? Lang Learn Technol. 2024;28(2):183–209 (. **Special Issue: Artificial Intelligence for Language Learning.**).
39. Han J, Yoo H, Kim Y, Myung J, Kim M, Lim H, Kim J, Lee TY, Hong H, Ahn S-Y. RECIPE: How to integrate ChatGPT into EFL writing education. In: Proceedings of the Tenth ACM Conference on Learning at Scale, pp. 416–420 (2023)
40. Lo CK, Yu PLH, Xu S, Ng DTK, Jong M-y. Exploring the application of ChatGPT in ESL/EFL education and related research issues: a systematic review of empirical studies. Smart Learn Environ. 2024;11(1):50.
41. Chen A, Zhang Y, Jia J, Liang M, Cha Y, Lim CP. A systematic review and meta-analysis of AI-enabled assessment in language learning: design, implementation, and effectiveness. J Comput Assist Learn. 2025;41(1):13064.

42. Mannuru NR, Shahriar S, Teel ZA, Wang T, Lund BD, Tijani S, et al. Artificial intelligence in developing countries: the impact of generative artificial intelligence (AI) technologies for development. Inf Dev. 2023;41(3):1036–54.
43. Liu L. Impact of AI gamification on EFL learning outcomes and nonlinear dynamic motivation: comparing adaptive learning paths, conversational agents, and storytelling. Educ Inf Technol. 2025;30(8):11299–338.
44. Petticrew M, Roberts H. Systematic reviews in the social sciences: a practical guide. Chichester, UK: John Wiley & Sons; 2008.
45. Cooper H. Research synthesis and meta-analysis: a step-by-step approach. 4th ed. Thousand Oaks, CA: Sage publications; 2015.
46. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372(71).
47. Boulton A, Cobb T. Corpus use in language learning: a meta-analysis. Lang Learn. 2017;67(2):348–93.
48. Plonsky L, Oswald FL. How big is "big"? Interpreting effect sizes in L2 research. Lang Learn. 2014;64(4):878–912.
49. Hedges LV, Olkin I. Statistical methods for meta-analysis. Orlando, FL: Academic Press; 1985.
50. Chen M-r. The AI chatbot interaction for semantic learning: a collaborative note-taking approach with EFL students. Lang Learn Technol. 2024;28(1):1–25.
51. Chen M-R. Improving English semantic learning outcomes through AI chatbot-based ARCS approach. Interact Learn Environ. 2025;33(6):3909–24.
52. Zheng Y-B, Zhou Y-X, Chen X-D, Ye X-D. The influence of large language models as collaborative dialogue partners on EFL English oral proficiency and foreign language anxiety. Comput Assist Lang Learn. 2025;1–27.
53. Zheng S. The effects of chatbot use on foreign language reading anxiety and reading performance among Chinese secondary school students. Comput Educ Artif Intell. 2024;7:100271.
54. Wiyaka W, Silitonga LM, Sunardi S, Pramudi YTC. From nervous to fluent: the impact of AI chatbot-assisted assessment on English reading anxiety and performance in Indonesia. Theory Pract Lang Stud. 2024;14(12):3851–60.
55. Wu Q, Xu A. Poe or Gemini for fostering writing skills in Japanese upper-intermediate learners: Uncovering the consequences on positive emotions, boredom to write, academic self-efficacy and writing development. Br Educ Res J. 2025;1–36.
56. Wang M, Akhter S. Tracing interpersonal emotion regulation, behavioural emotion regulation strategies, hopelessness and vocabulary retention within Bing vs. ChatGPT environments. Br Educ Res J. 2025;1–28.
57. Alsager H. To look from another window in education: artificial intelligence assisted language learning and its reflections on academic demotivation, foreign language learning anxiety and autonomy. Comput Assist Lang Learn Electron J. 2024;25(4):124–47.
58. Ghafouri M. ChatGPT: the catalyst for teacher-student rapport and grit development in L2 class. System. 2024;120:103209.
59. Ghafouri M, Hassaskhah J, Mahdavi-Zafarghandi A. From virtual assistant to writing mentor: exploring the impact of a ChatGPT-based writing instruction protocol on EFL teachers' self-efficacy and learners' writing skill. Lang Teach Res. 2024;1–23.
60. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. Introduction to meta-analysis. Chichester, UK: John Wiley & Sons; 2021.
61. Guo K, Pan M, Li Y, Lai C. Effects of an AI-supported approach to peer feedback on university EFL students' feedback quality and writing ability. Internet High Educ. 2024;63:100962.
62. Alnasib BN, Alharbi NS. Challenges and motivation: assessing Gemini's impact on undergraduate EFL students in classroom settings. World J English Lang. 2024;14(5):501–501.
63. Urazbayeva G, Kussainova R, Aibergen A, Kaliyeva A, Kantayeva G. Innovation off the bat: bridging the ChatGPT gap in digital competence among English as a foreign language teachers. Educ Sci. 2024;14(9):946.
64. Jaashan HMS, Alashabi AA. Using AI large language model (LLM-ChatGPT) to mitigate spelling errors of EFL learners. 2025;7(3):328–339.
65. Xu J, Liu Q. Uncurtaining windows of motivation, enjoyment, critical thinking, and autonomy in AI-integrated education: Duolingo vs. ChatGPT learning and motivation. Learn Motiv. 2025;89:102100.
66. Hwang W-Y, Luthfi MI, Liu Y-F. AI-enhanced video drama-making for improving writing and speaking skills of students learning English as a foreign language. Innov Lang Learn Teach. 2024;1–18.
67. Yılmaz ÖK, Aydın S. The impact of the use of artificial intelligence-generated materials on reading motivation among EFL learners. Read Res Q. 2025;60(3):70016.
68. Jamshed M, Albedah F, Ansari MS, Banu S. Assessing the efficacy of AI-driven corrective feedback via WhatsApp application to improve ESL learners' writing skills: an experimental study. Int J Interact Mob Technol. 2025;19(7):190–205.
69. Alwasidi MA, Al-Khalifah KS. Assessing the impact of ChatGPT on EFL students' writing productivity and proficiency. J Lang Teach Res. 2025;16(3):986–95.
70. Çelik F, Yangın Ersanlı C, Arslanbay G. Does AI simplification of authentic blog texts improve reading comprehension, inferencing, and anxiety? A one-shot intervention in Turkish EFL context. Int Rev Res Open Distrib Learn. 2024;25(3):287–303.
71. Yang G, Wang Y, Zhang Y, Yang M, Zeng Q, Song Z. An empirical study of AI-supported interleaved training strategy to improve EFL students' English impromptu speaking performance, learning engagement, technology acceptance and epistemic network structure. Asia Pac Educ Res. 2025. https://doi.org/10.1007/s40299-024-00962-y.
72. Kim R. Effects of learner uptake following automatic corrective recast from artificial intelligence chatbots on the learning of English caused-motion construction. Lang Learn Technol. 2024;28:109–33.
73. Ma Y, Chen M. AI-empowered applications effects on EFL learners' engagement in the classroom and academic procrastination. BMC Psychol. 2024;12(1):739.
74. Vicente-Yagüe-Jara M-I, López-Martínez O, Navarro-Navarro V, Cuéllar-Santiago F. Writing, creativity, and artificial intelligence: ChatGPT in the university context. Comunicar. 2023;31(77):45–54.
75. Oktarina IB, Saputri MEE, Magdalena B, Hastomo T, Maximilian A. Leveraging ChatGPT to enhance students' writing skills, engagement, and feedback literacy. Edelweiss Appl Sci Technol. 2024;8(4):2306–19.
76. Zahran FA. The impact of using Poe ChatGPT-based TPACK model on English as a foreign language teachers' performance and their students' vocabulary learning. High Learn Res Commun. 2025;15(1):1–11.

77. Teng MF, Huang J. Incorporating ChatGPT for EFL writing and its effects on writing engagement. International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT). 2025;15(1):1–21.
78. Wang C, Zou B, Du Y, Wang Z. The impact of different conversational generative AI chatbots on EFL learners: an analysis of willingness to communicate, foreign language speaking anxiety, and self-perceived communicative competence. System. 2024;127:103533.
79. Pan M, Lai C, Guo K. Effects of genai-empowered interactive support on university efl students' self-regulated strategy use and engagement in reading. Internet High Educ. 2025;65:100991.
80. Alsofyani AH, Barzanji AM. The effects of chatgpt-generated feedback on saudi efl learners' writing skills and perception at the tertiary level: a mixed-methods study. J Educ Comput Res. 2025;63(2):431–63.
81. Darmawansah D, Rachman D, Febiyani F, Hwang G-J. Chatgpt-supported collaborative argumentation: integrating collaboration script and argument mapping to enhance efl students' argumentation skills. Educ Inf Technol. 2025;30(3):3803–27.
82. Zheldibayeva R. GenAI as a learning buddy for non-English majors: effects on listening and writing performance. Educ Process Int J (EDUPIJ). 2025;14(1):1–16.
83. Kemelbekova Z, Degtyareva X, Yessenaman S, Ismailova D, Seidaliyeva G. AI in teaching English as a foreign language: effectiveness and prospects in Kazakh higher education. XLinguae. 2024;17(1):69–83.
84. Song C, Song Y. Enhancing academic writing skills and motivation: assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. Front Psychol. 2023;14:1260843.
85. Almineeai AM, Airashed MA, Hezam TA, Alharthi AF, Alharthi SF. The impact of ChatGPT on EFL medical track students' affective filter. J Lang Teach Res. 2025;16(1):158–67.
86. Liu W, Wang Y. The effects of using AI tools on critical thinking in English literature classes among EFL learners: an intervention study. Eur J Educ. 2024;59(4):12804.
87. Esfandiari R, Allaf-Akbary O. Assessing interactional metadiscourse in EFL writing through intelligent data-driven learning: the Microsoft Copilot in the spotlight. Lang Test Asia. 2024;14(1):51.
88. Boudouaia A, Mouas S, Kouider B. A study on ChatGPT-4 as an innovative approach to enhancing English as a foreign language writing learning. J Educ Comput Res. 2024;62(6):1509–37.
89. Sayed BT, Bani Younes ZB, Alkhayyat A, Adhamova I, Teferi H. To be with artificial intelligence in oral test or not to be: a probe into the traces of success in speaking skill, psychological well-being, autonomy, and academic buoyancy. Lang Test Asia. 2024;14(49):1–21.
90. Aldosari M. Another world with artificial intelligence in speaking classes: to delve into the influences on willingness to communicate, personal best goals, and academic enjoyment. Comput Assist Lang Learn Electron J. 2024;25(4):439–63.
91. Esfandiari R, Akbary OA. Assessing the impact of Microsoft Copilot and ChatGPT on EFL learners' interactional metadiscourse in argumentative writing. Innoeduca Int J Technol Educ Innov. 2025;11(1):47–73.
92. Sapan M, Uzun L. The effect of ChatGPT-integrated English teaching on high school EFL learners' writing skills and vocabulary development. Int J Educ Math Sci Technol. 2024;12(6):1679–99.
93. Meyer J, Jansen T, Schiller R, Liebenow LW, Steinbach M, Horbach A, et al. Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. Comput Educ Artif Intell. 2024;6:100199.
94. Kucuk T. ChatGPT integrated grammar teaching and learning in EFL classes: a study on Tishk International University students in Erbil, Iraq. Arab World Engl J. 2024. https://doi.org/10.24093/awej/chatgpt.6.
95. Wu T-T, Hapsari IP, Huang Y-M. Effects of incorporating AI chatbots into think–pair–share activities on EFL speaking anxiety, language enjoyment, and speaking performance. Comput Assist Lang Learn. 2025;1–39.
96. Liu Z-M, Hwang G-J, Chen C-Q, Chen X-D, Ye X-D. Integrating large language models into EFL writing instruction: effects on performance, self-regulated learning strategies, and motivation. Comput Assist Lang Learn. 2024;1–25.
97. Kim A, Su Y. How implementing an AI chatbot impacts Korean as a foreign language learners' willingness to communicate in Korean. System. 2024;122:103256.
98. Kim Y, Belcher D, Peyton C. Comparing monomodal traditional writing and digital multimodal composing in EAP classrooms: linguistic performance and writing development. J Engl Acad Purp. 2023;64:101247.
99. Zhou Q, Hashim H, Sulaiman NA. Supporting english speaking practice in higher education: the impact of AI chatbot-integrated mobile-assisted blended learning framework. Educ Inf Technol. 2025. https://doi.org/10.1007/s10639-025-13401-2.
100. Fathi J, Rahimi M, Derakhshan A. Improving EFL learners' speaking skills and willingness to communicate via artificial intelligence-mediated interactions. System. 2024;121:103254.
101. Rad HS, Alipour R, Jafarpour A. Using artificial intelligence to foster students' writing feedback literacy, engagement, and outcome: a case of Wordtune application. Interact Learn Environ. 2024;32(9):5020–40.
102. Rosenthal R. Meta-analysis: a review. Psychosom Med. 1991;53(3):247–71.
103. Sterne JA, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, Cates CJ, Cheng H-Y, Corbett MS, Eldridge SM et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019;366.
104. Ioannidis JP. Why most discovered true associations are inflated. Epidemiology. 2008;19(5):640–8.
105. Wang Y, Xue L. Using AI-driven chatbots to foster Chinese EFL students' academic engagement: an intervention study. Comput Hum Behav. 2024;159:108353.


## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.