The effectiveness of using artificial intelligence-assisted tools in second and foreign language learning: a systematic review

Yining Han

Department of Education, University of Oxford

Kellogg College

Supervisor: Dr. Elizabeth Wonnacott

Dissertation submitted in part-fulfilment of the requirements for the degree of

Master of Science in Applied Linguistics and Second Language Acquisition, 2024

DECLARATION BY THE CANDIDATE AS AUTHOR OF THE DISSERTATION

1. I understand that I am the owner of this dissertation and that the copyright rests with me unless I specifically transfer it to another person.
2. I allow the Department to deposit on my behalf a copy of this dissertation in the Oxford University Research Archive ('ORA') where it shall be freely available online for use in accordance with ORA's Terms and Conditions of Use [https://ora.ox.ac.uk/terms_of_use].
3. I understand that this dissertation should not contain material that can be used to personally identify individuals or specific groups of individuals (unless permission has been obtained from the individuals) and that such material should be removed before this dissertation is deposited in ORA.
4. I agree to be bound by the terms of the ORA Grant of Non-exclusive Licence [https://ora.ox.ac.uk/deposit_agreements] and I warrant that to the best of my knowledge, making my thesis available on the internet will not infringe copyright or any other rights of any other person or party, nor contain defamatory material.
5. I agree that my dissertation shall be available for download in ORA in accordance with paragraphs 2, 3 and 4 above.

Signed [an electronic signature is sufficient]:	Yining Han
Date:	08/08/2024

Plain language summary

Background

Recent years have witnessed significant progress in artificial intelligence (AI). Since the release of ChatGPT, a large language model developed by OpenAI, in 2022, AI has had a greater influence on various fields, including language education. AI technologies, such as intelligent tutoring systems and chatbots, have been increasingly integrated into language learning, demonstrating their potential to alleviate limitations on personalisation and interaction in traditional classrooms.

Objectives

This review includes studies published since 1956, when AI was first conceptualised, to investigate the current state and empirical evidence of the effectiveness of AI-assisted second or foreign language (L2/FL) learning. The review aims to summarise the characteristics of relevant studies and assess the impact of educational AI tools on language learning outcomes and learning perceptions.

Method

The review adopts a systematic review approach, retrieving 3547 records from nine electronic databases and citation searching. According to pre-defined eligibility criteria, the retrieved data were screened twice with the assistance of a second reviewer, resulting in the identification of 105 studies that met the inclusion criteria. These studies were first systematically mapped, followed by an in-depth review of those in higher education contexts to ensure the feasibility of data analysis within the specified time. Individual studies also went through quality assessment based on their study designs to ensure the reliability and validity of reported findings.

Results

Mapping the 105 eligible studies revealed that most research involved tertiary-level English as a second or foreign language (ESL/EFL) learners in Asian countries, focusing on five main types of AI-powered language learning tools (i.e., chatbots, AI-assisted language learning applications or platforms, AI-assisted writing and evaluation tools, intelligent tutoring systems, and speech recognition tools) and eight language skills (i.e., vocabulary, grammar, pronunciation, listening,

speaking, reading, writing, and integrated skills). Due to time constraints, the in-depth analysis only delved into 71 studies in tertiary settings and identified that 51 studies reported findings on learning outcomes, which showed generally consistent findings on enhanced outcomes, except for four studies. Additionally, the synthesis of 59 studies on learning perceptions revealed consistent findings of improved motivation and willingness to communicate but divergent results on general attitudes, learning anxiety, and learning interest. However, problems with study designs, such as an omission of confounding variables, a lack of transparent outcome measures, and an over-reliance on self-report data, might hinder the trustworthiness of findings on the effectiveness of AI-assisted language learning.

Implications

This review suggests several implications for future AI-assisted language learning research and practice. Firstly, the findings guide researchers to employ more robust study designs and explore less-investigated AI technologies in primary and secondary education settings. Secondly, language learners and teachers can use AI-driven tools to facilitate personalised and interactive learning of various language skills. Finally, the limitations of existing AI tools may inform developers of directions for future technological improvements.

Abstract

With the rapid evolution of artificial intelligence (AI), the potential of AI technologies to support language learning and teaching has captured increasing scholarly attention. This study systematically reviews publications since 1956 to provide an overview of the characteristics of existing research on AI-assisted second or foreign language (L2/FL) learning and assess its effect on learning outcomes and learning perceptions. The systematic review first presents a mapping of 105 eligible studies, identifying (1) a notable increase in the number of research outputs over the past four years; (2) a particular interest in tertiary-level language learners; (3) a primary focus on chatbots and writing skills as the most researched AI technology and language skill; and (4) a frequent absence of theoretical frameworks in research. The review then conducts an in-depth analysis of 71 studies in higher education contexts for the feasibility of data synthesis within the scheduled time. Findings revealed that (1) 92% of the 51 studies reporting learning outcomes demonstrated positive results; (2) learner perceptions of AI-assisted language learning were mixed, including consistent findings of enhanced motivation and willingness to communicate, but varied findings on learning attitudes, anxiety, and interest. However, the medium risk of bias across studies due to methodological flaws, such as lack of control groups and over-reliance on self-report data, suggested that the findings could not be taken as robust evidence for the effectiveness of AI-assisted language learning. The review provides the following theoretical and practical implications: (1) researchers should adopt more rigorous study designs, explore the application of AI technologies in primary and secondary education, and incorporate educational theories into future studies; (2) learners and teachers should keep utilising AI tools to facilitate personalised and interactive language learning; (3) developers should address the identified problems with AI-assisted language learning tools and promote further technological innovations.

Keywords: artificial intelligence, language learning, systematic review

Acknowledgements

First, I would like to extend my sincerest gratitude to my supervisor, Dr Elizabeth Wonnacott, for her invaluable expertise and feedback. Our supervision meetings were a tremendous source of inspiration, and the completion of this work could not have been possible without her support.

I am also grateful to my colleagues and friends at Oxford. A heartfelt thank you goes to Ivy Chan and Ivy Wang for their support for this project. A particular thanks also goes to Julius Han for his companionship during this journey.

Finally, thank you, Mum and Dad, for standing by me with steadfast encouragement and unwavering love.

Table of Contents

Plain language summary ..... **ii**
Abstract ..... **iv**
Acknowledgements ..... **v**
List of figures ..... **viii**
List of tables ..... **ix**
List of abbreviations ..... **x**
Chapter 1 Introduction **12**
1.1 Background to the study 12
1.2 Rationale and objectives of the study 12
1.3 Dissertation outline 13
Chapter 2 Literature Review **14**
2.1 Artificial intelligence 14
2.2 Artificial intelligence in education (AIEd) 15
2.3 Artificial intelligence in language learning 16
2.3.1 Theoretical underpinnings of AI-assisted language learning *16*
2.3.2 Machine translation *18*
2.3.3 Automatic speech recognition *19*
2.3.4 Automated writing evaluation *21*
2.3.5 Intelligent tutors *21*
2.3.6 Chatbot *23*
2.4 Previous reviews of AIEd 24
Chapter 3 Method **27**
3.1 Why a systematic review was chosen? 27
3.2 Protocol creation and registration 27
3.3 Eligibility criteria 28
3.4 Information sources 32
3.5 Search strategy 33
3.6 Data management 34
3.7 Data screening and selection process 34
3.7.1 Title and abstract screening *34*
3.7.2 Full-text screening *34*
3.7.3 Quality assessment *35*
3.8 Data extraction 36
3.9 Risk of bias assessment 37
3.10 Data synthesis 37
3.11 Confidence in cumulative evidence 38
Chapter 4 Results **39**
4.1 Mapping of included study characteristics 39
4.1.1 Publication trends *39*
4.1.2 Participant characteristics *40*
4.1.3 Geographical distribution *41*
4.1.4 Adopted AI technologies *42*
4.1.5 Studied language skills *43*
4.1.6 Types of study design *43*
4.1.7 Theoretical backgrounds *44*
4.1.8 Risk of bias assessment *44*
4.1.9 Keyword visualisation *45*
4.2 In-depth analysis of studies in higher education 46
4.2.1 Evidence of learning outcomes from vocabulary studies *47*
4.2.2 Evidence of learning outcomes from grammar studies *49*
4.2.3 Evidence of learning outcomes from pronunciation studies *50*

4.2.4 Evidence of learning outcomes from listening studies *50*
4.2.5 Evidence of learning outcomes from speaking studies *52*
4.2.6 Evidence of learning outcomes from reading studies *55*
4.2.7 Evidence of learning outcomes from writing studies *56*
4.2.8 Evidence of learning outcomes studies on integrated language skills *59*
4.3 Evidence of learning perceptions 61
4.3.1 General attitudes *62*
4.3.2 Motivation *63*
4.3.3 Willingness to communicate *64*
4.3.4 Learning anxiety *64*
4.3.5 Learning interest *65*
Chapter 5 Discussion 67
5.1 RQ1: What is the extent of research that has been conducted? 67
5.2 RQ2: What is the extent of empirical evidence for the effectiveness of AI-assisted language learning? 69
5.2.1 The extent of findings on learning outcomes *69*
5.2.2 The extent of findings on learning perceptions *72*
6 Conclusion 73
6.1 Summary of findings 73
6.2 Theoretical and practical implications 73
6.3 Limitations and future research 74
References 75
Appendix 1. The review protocol 91
Appendix 2. Sample Boolean search strings 104
Appendix 3. The data extraction form 105
Appendix 4. MMAT and explanations 109
Appendix 5. Detailed information for each study 122
Appendix 6. Risk of bias of individual studies 130

List of figures

Figure 1. The relationship between AI, ML, and DL 14
Figure 2. Flow diagram of the data screening and selection process (adapted from Page et al., 2020) 36
Figure 3. Publication trends 39
Figure 4. Participants' educational levels 40
Figure 5. Participants' L1 and L2/FL 41
Figure 6. Participants' L2/FL proficiency 41
Figure 7. Location of study 42
Figure 8. Types of AI technologies 42
Figure 9. Target language skills 43
Figure 10. Types of study design 43
Figure 11. Theoretical groundings of included studies 44
Figure 12. Keyword mapping 45
Figure 13. Reported learning outcomes and RoB 70

List of tables

Table 1. Eligibility criteria 28
Table 2. A list of consulted databases 32
Table 3. Definition of learning outcome categories 47
Table 4. Extracted data from vocabulary studies 48
Table 5. Extracted data from grammar studies 49
Table 6. Extracted data from pronunciation studies 50
Table 7. Extracted data from listening studies 51
Table 8. Extracted data from speaking studies 53
Table 9. Extracted data from reading studies 55
Table 10. Extracted data from writing studies on chatbots 56
Table 11. Extracted data from writing studies on AI-assisted writing and evaluation tools 57
Table 12. Extracted data from studies on integrated language skills 60

List of abbreviations

AI = artificial intelligence

AIEd = AI in education

AILEd = AI in language education

ASR = automatic speech recognition

AWE = automated writing evaluation

CG = control group

CNN = Convolutional Neural Network

DL = deep learning

EFL = English as a foreign language

EG = experimental group

ESL = English as a second language

FL = foreign language

GAN = Generative Adversarial Network

GenAI = generative AI

GPT = Generative Pre-trained Transformer

GRADE = Grading of Recommendations Assessment, Development, and Evaluation

IDESR = International Database of Education Systematic Reviews

ITS = intelligent tutoring system

L1 = first language

L2 = second language

LLM = large language model

ML = machine learning

MMAT = Mixed Methods Appraisal Tool

MT = machine translation

NLP = natural language processing

PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analysis

RCT = randomised controlled trial

RNN = Recurrent Neural Network

RoB = risk of bias

RQ = research question

SCT = Sociocultural Theory

SLA = second language acquisition

SOLO = Search Oxford Libraries Online

ZPD = Zone of Proximity Development

Chapter 1 Introduction

1.1 Background to the study

Artificial intelligence (AI) technologies have revolutionised the educational sector (Holmes et al., 2019; Luckin et al., 2016; UNESCO, 2023). Rapid advancements in AI techniques, such as machine learning (ML), deep learning (DL), and natural language processing, have provided momentum for the implementation of AI in education (AIEd) (Ouyang et al., 2022; Zawacki-Richter et al., 2019). Consequently, AI technologies and their applications, such as chatbots, intelligent tutoring systems, and automated scoring systems, have been increasingly adopted in the learning and teaching process (X. Chen et al. 2020; X. Chen et al., 2022). With the launch of ChatGPT, a generative AI chatbot powered by large language models (LLMs), in late 2022, AI-assisted education has garnered ever-greater scholarly attention (Han, 2024; Wang et al., 2024). The advent of the generative AI (GenAI) era has ushered in new opportunities and increasing investments for the realm of education. According to statistics provided by UNESCO (2021), the global AIEd market size is estimated to reach $6 billion by 2024.

Numerous AI-powered educational tools have been utilised specifically for language learning, teaching, and assessment (X. Chen et al., 2022). For example, Duolingo, a popular language learning application with over 500 million users, employs ML, DL, and LLMs to provide personalised learning paths and interactive learning opportunities (Qiao & Zhao, 2023). Additionally, iFlyTek, a leading speech recognition company, have developed intelligent tutoring systems to support the English Speaking Test in China's National College Entrance Examination (iFlyTek, n.d.). The success of these applications highlights the significant potential of integrating AI technologies into language education.

1.2 Rationale and objectives of the study

In second or foreign language (L2/FL) education, conventional language classrooms tend to be restricted by insufficient practice opportunities (Fathi et al., 2024), lack of real-time feedback (Wu et al., 2017), and limited personalisation (Luo, 2016). Empirical studies have demonstrated the potential of AI-supported educational tools in overcoming these limitations, showing their capabilities to provide supplementary learning materials (Wei, 2023), offer immediate corrective feedback (Fu et al., 2020), enable personalised learning (Zhou, 2023), and enhance learning outcomes (Fathi et al., 2024).

While numerous studies have documented the effectiveness of AI-powered tools in language learning, systematic reviews in this specific field remain scarce, especially when compared to broader reviews of AIEd (Huang et al., 2023; Law, 2024). A further research gap is that few reviews have included studies on GenAI and LLMs, which might be insufficient for understanding the latest research trends in adopting these emerging technologies for language learning. As suggested by Han (2024), systematic research was needed to evaluate the role of GenAI tools represented by ChatGPT. To address these gaps, this review aims to include studies dating from 1956, when the concept of AI was proposed, to more recent studies on GenAI. It seeks to analyse and synthesise studies on the effectiveness of AI-driven tools in L2/FL learning. The findings are expected to identify the current state of research and inform future language learning research and practice.

1.3 Dissertation outline

The dissertation has six chapters. This chapter introduces the research background and objectives. Chapter 2 reviews existing literature on AIEd and AI-assisted language learning and presents the research questions. Chapter 3 outlines the review methodology, followed by results and discussion in Chapter 4 and Chapter 5. Finally, Chapter 6 summarises the findings and provides theoretical and practical implications for language researchers, learners, and educators.

Chapter 2 Literature Review

This chapter is divided into four sections. It begins by defining AI and its related concepts, including machine learning and deep learning. The subsequent section discusses how AI can be applied in educational settings. The third section delves into the application of AI in language learning, reviewing its theoretical foundations and empirical studies. Finally, existing review studies on AI in language education are scrutinised to identify the research gap for this review.

2.1 Artificial intelligence

Artificial intelligence is an evolving and interdisciplinary research field that has driven scholars to propose multiple definitions (Russell & Norvig, 2010; Wang, 2019). This paper draws from McCarthy's (2007) definition and views AI as the ability of intelligent machines to simulate human intelligence while performing tasks and solving problems. According to Russell and Norvig (2016), AI is an overarching term for a range of subfields including machine learning (ML), deep learning (DL), and applications powered by these algorithms. This perspective coincides with that of scholars (cf. Sarker, 2021a; Sze et al., 2017; Webb et al., 2021), who have described DL as a subset of ML, which is a further subset of AI (see Figure 1).

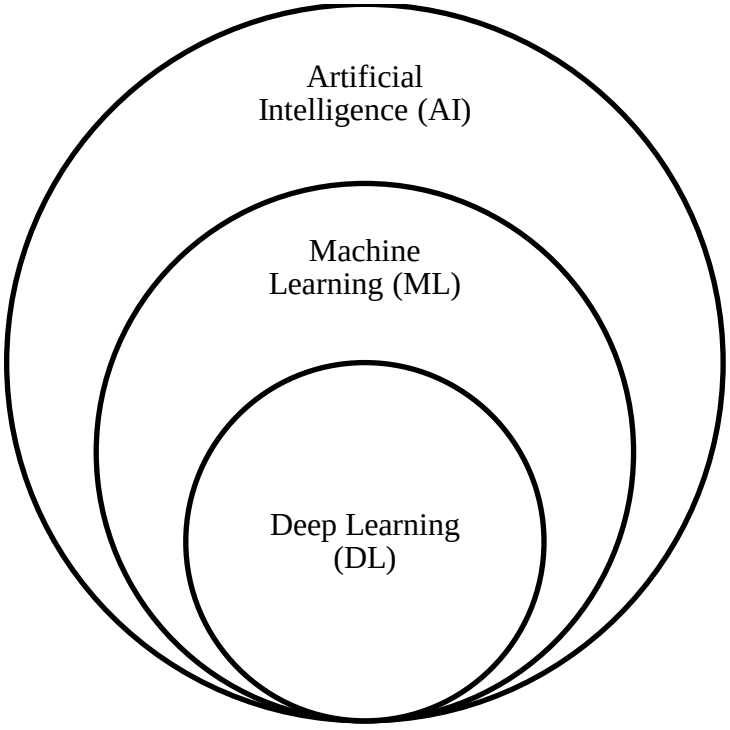


Figure 1. The relationship between AI, ML, and DL

Coined by Arthur Samuel (1959), *machine learning* refers to the capability of intelligent machines to learn from experience or training data without being explicitly programmed for specific tasks. Through the learning or training process, intelligent machines can recognise patterns, make predictions, and automatically address novel problems (Holmes et al., 2019; Popenici & Kerr, 2017; Sarker, 2021b).

Deep learning was introduced in 1986 as a subset of machine learning, which involves the ability of intelligent machines to generalise patterns from vast amounts of data through multi-layered processing (Sze et al., 2017). The recognition of data patterns is achieved by *neural networks* (also known as *artificial neural networks*), which consist of interconnected nodes that resemble neurons in human brains. Common types of neural networks include *Convolutional Neural Networks* (CNNs) that primarily model visual data, *Recurrent Neural Networks* (RNNs) that handle sequential data, and *Generative Adversarial Networks* (GANs) that generate videos and graphs (Gillani et al., 2023). Recent breakthroughs in neural network architectures, particularly with the introduction of the Transformer algorithm, have led to the development of large language models like the Generative Pre-trained Transformer (GPT) series (OpenAI, 2015). Specifically, the Transformer algorithm relies on the self-attention mechanism to learn more comprehensive relationships between sequential data compared to traditional sequential modelling approaches such as RNNs (Vaswani et al., 2017). As a result, the Transformer model possesses more robust text comprehension and generation capabilities. Built upon the Transformer architecture, GPT-series models are trained on large-scale datasets from the Internet, which are better at understanding human instructions and generating comprehensible and contextually appropriate outputs (Brown et al., 2020; Yenduri et al., 2024). Since the debut of GPT-1 in 2018, the GPT series has evolved through several iterations, including GPT-2, GPT-3, GPT-3.5 (also known as ChatGPT), GPT-4, and GPT-4o (OpenAI, 2024).

ML and DL algorithms have been utilised across different fields (LeCun et al., 2015; Webb et al., 2021). The following section will focus on how these AI technologies have been implemented in the educational sector.

2.2 Artificial intelligence in education (AIEd)

Significant advancements in AI technologies have contributed to their proliferation in diverse domains, with education being no exception (Gillani et al., 2023; Holmes et al., 2019; Luckin et al., 2016; Popenici & Kerr, 2017). Baker and Smith (2019) classified AIEd tools into three categories: (1) learner-facing tools, (2) teacher-facing tools, and (3) system-facing tools. *Learner-facing tools*, such as intelligent tutoring systems, are designed to meet students' personalised learning needs. *Teacher-facing tools*, such as automated scoring systems, are used to assist teaching activities.

System-facing tools aim to support school administrative work, such as timetable management and attendance tracking. Of note, this broad classification only provides a simplified overview of AI-supported educational tools. Applications in authentic pedagogical settings may combine functions from multiple categories. In summary, AIEd tools have great potential in (1) individualising students' learning paths, (2) improving teachers' working efficiency, and (3) facilitating administrators' tasks at school.

The impact of AIEd tools on learners, instructors, and educational administrators has long attracted scholarly interest. Extensive research has examined the application of AI in assisting students to learn a range of disciplines, such as art (Chiu et al., 2024), business (Montalvo et al., 2018), languages (Karataş et al., 2024), medicine (W. Zhang et al., 2024), and STEM subjects (Ouyang et al., 2023). Additionally, studies have explored how AI technologies could support pedagogical activities such as assessment (Hooshyar et al., 2016), grading (Zhao et al., 2022), and performance prediction (Gaudioso et al., 2012). Beyond the potential benefits of AIEd applications in learning and teaching, some scholars have raised concerns about over-reliance on AIEd tools and ethical issues such as data privacy (Bond et al., 2024; Gillani et al., 2023). Given the affordances and challenges, AIEd tools should be used meticulously to optimise their efficacy in educational contexts.

Several systematic reviews of AIEd have identified language education as a field that has gained particularly considerable attention (X. Chen et al., 2022; Hwang & Chang, 2023; Xie et al., 2019). This review intends to specifically concentrate on AIEd tools that incorporate functionalities facing language learners. The section below will examine empirical studies on AI-assisted language learning.

2.3 Artificial intelligence in language learning

This section elaborates on the use of AI-driven tools in language education, with a specific focus on how they support L2/FL learners. It begins by using theoretical frameworks within the field of second language acquisition (SLA) to discuss how AI tools can assist language learning. It then introduces five practical applications of AI-driven language learning.

2.3.1 Theoretical underpinnings of AI-assisted language learning

Three SLA theories can inform the implementation of AIEd tools in L2/FL learning, namely the Input Hypothesis (Krashen, 1981, 1982, 1985), the Interaction Hypothesis (Long, 1981, 1985, 1996), and the Sociocultural Theory (Vygotsky, 1978; 1986).

According to the Input Hypothesis (Krashen, 1981, 1982, 1985), language learning occurs naturally when learners are exposed to comprehensible input that slightly surpasses their current proficiency levels, denoted as *i* +1. Input that is overly simplistic fails to provide new linguistic structures and vocabulary, which may obstruct learners from progressing towards higher levels of competence. Conversely, excessively challenging input may limit learners' meaningful interaction with the language and increase their anxiety, inducing high affective filters and impeding the language acquisition process (Krashen, 1982). Therefore, educators should deliver content that aligns with learners' *i* + 1 levels. However, due to individual differences in linguistic levels, it would be virtually infeasible to provide optimal *i* + 1 input for each student in conventional language classrooms. The challenge may be resolved by AIEd tools. Applications such as intelligent tutoring systems can dynamically assess students and tailor teaching contents to students' competence levels, transforming the conventional one-size-fits-all educational approach into a personalised and individualised language acquisition experience. Such transformation could potentially enhance the comprehensibility of learning contents and address the affective filter, thereby optimising language learning outcomes.

Another theory that is useful in understanding how AI technologies may be integrated into language learning is the Interaction Hypothesis (Long, 1981, 1985, 1996), which posits that language acquisition is driven by meaningful interaction between interlocutors. When speakers detect communication breakdown in their conversations, they initiate negotiation by making interactional adjustments, which allows them to receive corrective feedback and produce modified output in the interaction (Long, 1996). According to this theory, the process of negotiation for meaning not only enhances mutual comprehensibility but also facilitates language acquisition (Gass, 1997). While the high student-to-teacher ratio in traditional classrooms limits interactions between teachers and students (Fathi et al., 2024), AIEd tools could potentially simulate the role of interlocutors and provide immediate feedback to help learners identify and correct their errors, thereby facilitating their language development.

Sociocultural Theory (SCT) (Vygotsky, 1978; 1986) also offers a valuable perspective for understanding the potential of AI in language learning. The theory highlights the role of social interaction in creating collaborative and authentic language learning contexts. A key construct in SCT is the Zone of Proximal Development (ZPD), which represents the disparity between what learners can achieve independently and what they can accomplish with the assistance of more knowledgeable individuals (Vygotsky, 1978). External support from more proficient others acts as scaffolding, enabling language learners to make progress within their ZPD (Vygotsky, 1986). In AI-assisted language learning, tools such as chatbots and intelligent tutoring systems potentially create interactive

learning environments for meaningful interaction with peers, teachers, and AI partners. Additionally, personalised feedback from AIEd tools can support learner's language development within their ZPD.

Grounded in these SLA theories, AI technologies hold promise for addressing some of the limitations of traditional classrooms by providing a more individualised and interactive learning environment. This review is motivated by the potential benefits of AIEd tools, which intends to assess the extent to which these benefits can be supported by empirical evidence. Before a systematic review of existing literature, the remaining parts of this section will examine how AI-assisted language learning applications may support L2/FL learners in various aspects of language skills.

2.3.2 Machine translation

Machine translation (MT), a common application of AI in language learning, utilises computer programs to translate texts from a source language to a target language (Jiang & Lu, 2020; Son et al., 2023). This process is powered by natural language processing (NLP), an AI technology that enables computers to recognise and comprehend human language through various techniques, including ML and DL algorithms (Russell & Norvig, 2010).

MT has undergone three stages of development: (1) rule-based MT, (2) statistical MT, and (3) neural MT (C. Zhang et al., 2023). Early *rule-based MT systems* relied on manually crafted linguistic rules for the source and target languages. These systems were time-consuming to construct, and their accuracy depended heavily on the quality of input rules. The advent of *statistical MT systems* marked a shift towards a data-driven approach, where systems were trained on bilingual corpora to identify statistical patterns in parallel texts and generate translations. More recently, sophisticated DL algorithms, such as the Transformer models, have been incorporated into *neural MT systems* represented by Google Translate (Jiang & Lu, 2020; Koehn, 2020; Lee, 2023). It is claimed that these advancements in DL could improve machine translator's capability to select contextually appropriate vocabulary, thereby tremendously enhancing translation accuracy and comprehensibility. Due to its potentially augmented performance, MT has been increasingly implemented in language education and has attracted greater empirical attention (Briggs, 2018).

One strand of studies has explored the effectiveness of MT in L2/FL education. Lee (2023) systematically reviewed 87 studies and reported a generally positive effect, particularly on enhancing writing outcomes. For example, Garcia and Pena (2011) examined beginning and pre-intermediate Spanish learners and divided them into two groups: one completed writing tasks directly in L2, while the other wrote in their first language (L1) and then edited the MT output. It was found that MT enabled both beginners and pre-intermediate learners to produce longer texts with fewer errors and a richer variety of vocabulary and tenses, reflecting an enhanced quantity and quality of writing output.

For advanced learners, Niño (2008) observed that revising MT outputs exerted a medium effect size on reducing lexical (d = -0.479) and grammatical errors (d = -0.443), and a large effect size on reducing spelling mistakes (d = -1.216), thereby improving the quality of translated texts. In both studies, the process of refining MT outputs resembles the negotiation of meaning in Long's (1996) Interaction Hypothesis. When learners encountered inappropriate translations, they rectified them to be comprehensible, which enhanced their linguistic knowledge. Accordingly, MT's capability to facilitate timely interactional adjustments can serve as an alternative to synchronous feedback from teachers, which is often limited in traditional classrooms.

However, improved writing output may not necessarily imply substantive gains in language skills. As Garcia and Pena (2011) noted, using MT might not facilitate language acquisition to the same extent as writing directly in an L2. Similarly, Karnal and Pereira (2015) cautioned in their reading study that while MT enhanced reading comprehension, it may not reliably facilitate vocabulary acquisition. This discrepancy between improved performance and actual language acquisition prompts further investigations into the pedagogical role of MT in language classrooms.

Another strand of research has thus examined students' perceptions of MT, yielding mixed results. While Niño (2009) reported that MT could boost learners' confidence in language learning, Valijärvi and Tarsoly (2019) found in their survey study that learners raised concerns that relying on MT would demotivate them from FL practices. Furthermore, learners expressed dissatisfaction with MT errors (Tsai, 2019), which may pose a significant challenge to beginners who lack the linguistic knowledge to critically assess the MT output. While language can be acquired through meaningful input and corrective feedback (Long, 1996), the provision of inaccurate feedback from MT may lead to the internalisation of incorrect linguistic forms, thereby hindering language acquisition.

Recent advancements in the Transformer algorithm are purported to improve translation quality. It is worthy of investigation whether these advancements result in different language learning outcomes and perceptions compared to previous studies. A systematic review is thereby needed to answer the query. Additionally, the introduction of the Transformer algorithm has revolutionised NLP, contributing to its expanding application in various language learning systems, programs, and applications (Yenduri et al., 2024). These applications will be specified in the sections below.

2.3.3 Automatic speech recognition

Automatic speech recognition (ASR) is an NLP-powered technology that enables computers to decode and transcribe spoken language into written texts (Shadiev & Liu, 2023). ASR systems can be categorised into (1) *speaker-dependent ASR systems*, which are trained on an individual speaker's speech samples and tailored to the speaker's voice features; (2) *speaker-independent ASR systems*,

which are trained on large-scale datasets to recognise general features of human speech; and (3) *speaker-adaptive ASR systems*, which are initially constructed based on large datasets and can be tailored to individual users over time (Young & Mihailidis, 2010). The second type of ASR system has been widely adopted in language learning tools especially speaking practice applications to provide interactive and personalised learning experiences.

Several empirical studies have substantiated the benefits of ASR in improving L2/FL speaking skills. Shafiee Rad (2024) evaluated the effect of an ASR-assisted speaking practice application on English speaking proficiency among Iranian university students. In this study, the experimental group (EG) utilised the application for speaking practice and feedback, whereas the control group (CG) completed traditional speaking homework and received feedback from teachers. Following a 10-week intervention, the comparison between pretest and post-test results revealed that the EG achieved significantly higher speaking scores than the CG, indicating a moderate effect of the application on L2 speaking proficiency ($\eta_p^2 = 0.177$). The improvement may be attributed to the instant corrective feedback on speech from the application, which allows learners to immediately modify their output, demonstrating the role of interactional adjustments in language learning (Long, 1996). Evers and Chen (2022) investigated the effect of an ASR dictation system on two dimensions of pronunciation, accentedness (i.e., the extent to which a speech deviates from the target variety) and comprehensibility (i.e., how easy to understand a speech), among Taiwanese adult English learners. It was found that the ASR system exerted a greater impact on improving L2 speech comprehensibility than on correcting accentedness. The outcome may reflect the inherent limitations of speaker-independent ASR systems, which are typically less sensitive to different accents. The inadequate corrective feedback on accents resulted in fewer opportunities for learners to identify and correct their pronunciation errors, thus impeding them from engaging in interactional adjustments and improving in this aspect of speaking.

Studies have also explored user perceptions of ASR, showing mixed findings. Evers and Chen (2022) noted that the participants rated the ASR system as satisfactory and easy to use. However, ASR sometimes frustrates users due to errors in speech recognition, especially when recognising non-native accents (McCrocklin, 2016). Technological advancements in neural network architectures, such as the self-attention mechanism, enable ASR systems to adapt to unfamiliar accents and enhance speech recognition accuracy by evaluating the relevance of different words and considering speech contexts. It is thus worthwhile to review the latest literature and examine whether these advancements have effectively addressed deficiencies in ASR.

2.3.4 Automated writing evaluation

Another application of NLP in language learning is automated writing evaluation (AWE), which is designed to provide quantitative scores and qualitative feedback on written assignments (Shermis et al., 2013). The implementation of AWE systems in L2/FL writing can be grounded in the Interaction Hypothesis (Long, 1996). The hypothesis posits that interactive adjustments in the negotiation of meaning process facilitate language acquisition. AWE systems provide instant feedback and enable learners to revise and resubmit their writings for multiple times. This interaction process allows students to continuously modify their writing and negotiate for more comprehensible meaning. Additionally, from a sociocultural perspective (Vygotsky, 1986), the personalised AWE feedback bridges the gap between students' current and potential writing performance, providing scaffolding within their ZPD to produce better writing.

Apart from theoretical underpinning, empirical studies have also investigated the effectiveness of AWE feedback in L2/FL writing. For example, Chang et al. (2021) examined the influence of AWE feedback on Chinese English as a foreign language (EFL) learners' writing performance. Participants with equivalent writing abilities completed five essays and received feedback from either Grammarly (EG) or their peers (CG) over 16 weeks. The post-test scores revealed that the EG performed significantly better than the CG, yielding a medium effect of AWE on FL writing performance (Cohen's $d = 0.603$). In another study, Link et al. (2022) divided Iranian EFL learners into an EG who received both AWE and teacher feedback and a CG who only received teacher feedback. Findings indicated that the EG showed significant improvement only in writing accuracy, whereas the CG improved significantly in complexity, accuracy, and fluency. This reflects the limitation of AWE, which typically provides linguistic-level feedback on grammar but lacks discursive-level feedback on logical flow. Comparatively, teacher feedback tends to be more comprehensive, identifying both grammatical mistakes and problems with coherence between sentences. Overall, AWE can serve as either an alternative or a complement to human feedback in L2/FL writing learning. With advancements in ML and NLP, AI-generated automated writing feedback is expected to consider the broad contexts in writing. Examining recent empirical studies can determine whether this trend can be validated.

2.3.5 Intelligent tutors

Intelligent tutors are instructional systems or software that also utilise AI techniques, such as ML and NLP, to provide adaptive and personalised learning (Mousavinasab et al., 2021; Shute & Psotka, 1996). The classical architecture of an intelligent tutoring system (ITS) consists of four components

(Corbett et al., 1997; Mousavinasab et al., 2021). The *domain model* contains the expert knowledge that the ITS intends to teach, encompassing a network of concepts, rules, and problem-solving techniques relevant to the domain. The second component is the *learner model*, which tracks user behaviours (e.g., task performance, learning pace, learning styles) and collects other relevant information (e.g., linguistic proficiency, learning goals, study preferences) to diagnose learners' needs and customise instructional paths. Based on data from the learner model, the ITS can identify knowledge that learners do not possess and offer adaptive pedagogical action in the *tutoring model*. The final constituent is the *user interface*, which is concerned with the environment where the teaching contents are delivered to students. Collectively, the interaction between the four components contributes to the functionality of an ITS to capture individual differences and tailor teaching content to learners' $i + 1$ levels (Krashen, 1982).

The first ITS, the SCHOLAR system, was created by Carbonell (1970) to answer students' questions and assess their knowledge of geography. The launch of SCHOLAR was followed by the development and application of numerous ITSs to support learning in multiple disciplines over the past few decades. In the field of language education, ITS applications have been introduced to assist the acquisition of diverse language skills, including vocabulary (Esit, 2011), grammar (Chia & Li, 2020), listening (Amaral et al., 2011), reading (Teo, 2012), speaking (Qian et al., 2023), and writing (Roscoe & McNamara, 2013). Specifically, ITSs for speaking and writing tutoring are enhanced by the ASR and AWE technologies introduced in the preceding sections. The implementation of ITSs for language learning can be informed by the interactionist approach (Long, 1996). By interacting with the ITSs, language learners notice the deficiencies in their knowledge and develop their language skills through corrective feedback from the systems (Long, 1996).

Empirical studies have investigated the effectiveness of ITSs and user perceptions. Esit (2011) designed an intelligent language learning system equipped with a morphological analyser. The system could divide morphologically complex vocabulary into its root and affixes and provide word use examples, which was expected to support English vocabulary learning and reading comprehension. The researcher divided Turkish EFL learners into an EG who used the system and a CG using dictionaries for reading activities. Following a six-week intervention, both groups improved their vocabulary test scores, and the increase was significantly more pronounced in the EG. Additionally, the EG reported significantly higher levels of perceived usefulness and lower levels of anxiety regarding the system. Focusing on Japanese grammar learning, Chia and Li (2020) developed an ITS featuring error identification and corrective feedback. The researchers examined the efficacy of the system among Taiwanese secondary school students, finding that the EG using the program improved significantly more than the CG receiving teacher-led instruction in the grammar test after a 40-minute intervention. Moreover, in Xu et al.'s (2019) meta-analysis, the 19 included studies yielded a large

effect size (ES = 0.86) on the effectiveness of ITSs in improving K-12 students 'reading comprehension compared to traditional classroom instructions.

Altogether, studies have adopted pretest-posttest control designs to compare ITS-assisted instruction against traditional teacher-led instruction, showing consistently positive effects of ITSs and generally favourable user perceptions. However, one limitation of existing ITSs is their narrow focus on individual linguistic skills. Investigations into advanced ITSs that combine training for integrated skills are needed to explore their efficacy in developing overall language proficiency. Another area where AI demonstrates significant potential in assisting students in real time is the use of chatbots, which will be specified in the next section.

2.3.6 Chatbot

Chatbot, also known as chat bot or conversational agent, is a computer program combining a range of AI techniques, including ML, NLP, and ASR, to enable communication between machines and humans via texts or audio (Adamopoulou & Moussiades, 2020). Chatbots have their origin in ELIZA, which was invented by Weizenbaum (1966) at the Massachusetts Institute of Technology to simulate a psychotherapist and answer users' queries based on pattern matching. The core working mechanism of ELIZA involved identifying keywords and phrases in user inputs and then selecting pre-programmed rules and scripts as responses. However, due to its rule-based system and limited knowledge base, ELIZA could not learn and could only engage in short, domain-specific conversations with users.

Since the invention of ELIZA, chatbots have progressed constantly in the subsequent decades, driven by advancements in AI technologies (Adamopoulou & Moussiades, 2020). Modern chatbots incorporate ML algorithms and sentiment analysis to learn from large datasets and generate more natural and contextually appropriate responses. Additionally, the utilisation of ASR technology in voice assistants such as Apple Siri, Google Assistant, and Amazon Alexa allows oral interactions between machines and humans. A notable milestone in the evolution of chatbots is the creation of ChatGPT (https://chatgpt.com/) by OpenAI in November 2022 based on the GPT architecture. ChatGPT is trained on vast amounts of data and can be fine-tuned for specific NLP tasks such as answering questions and generating texts (OpenAI, 2022). The fine-tuning process allows ChatGPT to learn a range of language patterns and produce human-like texts. Recent iterations of LLMs, including GPT-4 and GPT-4o, have been released to further enhance ChatGPT's text generation capabilities.

Chatbots have been applied in various areas, including language education (Hwang & Chang, 2023). Since the introduction of ChatGPT, chatbot-assisted L2/FL learning has garnered increasing

empirical attention (Han, 2024). Studies have focused on various aspects of language learning, including vocabulary (Zhang & Huang, 2024), grammar (Kim, 2019), listening (Kim, 2018), reading (Liu et al., 2022), speaking (Yuan, 2023), and writing (Kwon et al., 2023).

Empirical studies have demonstrated the effectiveness of chatbots in enhancing language learning outcomes through a pretest-posttest research design (Kim, 2018; Kim, 2019; Kwon et al., 2023; Liu et al., 2022; Yuan, 2023; Zhang & Huang, 2024). Comparisons between chatbot-assisted EGs and CGs without chatbot assistance have consistently shown that the EG exhibited significantly greater improvements in the post-test scores. Meta-analyses have further corroborated an overall positive effect of chatbots on language learning. For example, Zhang and colleagues (2023a) synthesised 18 studies on chatbot-assisted language learning, identifying an overall moderate effect ($g = 0.527$).

However, studies on user perceptions of chatbots have reported mixed findings. Some studies have identified benefits of chatbots in language learning, including (1) creating interactive and authentic learning environments (Jeon, 2021), (2) offering abundant practice opportunities (Kim, 2018), and (3) reducing language learners' anxiety (Ayedoun et al., 2019). These benefits can be supported by the interaction theory, which highlights the role of interaction in language acquisition. Nevertheless, mixed affective learning outcomes have been yielded. Liu et al. (2022) observed that Taiwanese elementary school students maintained reading interest after having a book talk with a chatbot while reading. By contrast, Fryer et al. (2017) asked Japanese EFL students to engage in tasks with human and chatbot partners, finding that students were initially interested in chatbots due to the "novelty effect" of being exposed to new technologies (Fryer et al., 2017, p. 463), but this interest declined as time went on. Additionally, chatbots sometimes generate inaccurate and irrelevant responses, leading to unsatisfactory user experiences (Coniam, 2014).

With advancements in AI technology, particularly the Transformer model, some limitations of chatbots are likely to be resolved. A comprehensive review of AI-powered language learning tools is needed to identify the latest technological developments and research trends. The following section will examine extant review studies on AI-assisted education to uncover research gaps.

2.4 Previous reviews of AIEd

Driven by the increasing application of AIEd tools, a number of reviews have been conducted to synthesise research findings. Zawacki-Richter et al. (2019) conducted a systematic review of 146 research papers on AI in higher education and identified four roles of AIEd tools, including (1) profiling and predicting students' academic performance, (2) providing automated assessment and evaluation, (3) recommending adaptive and personalised learning materials, and (4) offering

customised tutoring. X. Chen et al. (2020) systematically reviewed 45 AIEd studies and concluded that there was a growing research interest, particularly in NLP-powered AIEd applications such as ITSs. However, a notable lack of educational theories in these studies suggested a need to integrate pedagogical theories into future research. Wang et al. (2024) included 125 studies in their systematic review, identifying similar roles of AIEd tools to those in Zawacki-Richter et al. (2019) and noting a rise in studies on emerging technologies like educational chatbots driven by advanced DL algorithms. The review also echoed X. Chen et al. (2020) in finding a lack of theoretical underpinnings in most studies. In one of the few meta-analyses, Zheng et al. (2023) reviewed (quasi-)experimental studies on the impact of AI on learning achievement and learning perception. All included studies compared EGs who adopted AI-assisted learning against CGs who received traditional instruction without AI assistance. The synthesis of 24 eligible publications indicated an overall high effect size of AI on learning achievement (ES= 0.812) and a small effect size on learning perception (ES= 0.208), implying that using AI generally helped participants gain better learning performance and more positive learning experiences. In sum, these studies provided an overview of the research trends and potential of AIEd.

Compared to reviews of AI in general education, relatively fewer review studies have focused specifically on AI in language education (AILEd). For instance, Yang and Kyun (2022) conducted a systematic review of 25 research papers on AI-assisted language learning published between 2007 and 2010, identifying publication trends in the use of ITSs and automated evaluation systems to improve various language skills, particularly writing skills. However, one limitation of the review was its collection of literature over a narrow timeframe and from merely two databases, resulting in limited generalisability of findings. Similarly, Huang et al. (2023) reviewed 516 papers published from 2000 to 2019, reporting research foci on using AWE systems and ITSs for writing and reading. Despite the substantial number of included articles, the review also suffered from a small number of databases (n = 3) and a short time span, rendering it insufficient to provide comprehensive findings on trending research topics.

To address the temporal limitations in preceding reviews, Liang and colleagues (2023) set a longer period for including studies. Their systematic review of 71 AILEd publications between 1990 and 2020 identified the following research trends: (1) the past decade witnessed a notable surge in publications, with particular research interests in ITS and NLP systems for writing, reading, and vocabulary acquisition; (2) AILEd's primary roles include serving as intelligent tutors to guide language learning, functioning as intelligent evaluators to assess learning outcomes, and providing personalised language learning plans; (3) the majority of studies focused on the cognitive domain of learning outcomes (e.g., substantive language acquisition), followed by those reporting affective (e.g.,

motivation) and behavioural outcomes (classroom interaction). However, the review was also limited by collecting publications from one database, rendering it still not comprehensive enough.

Taken together, a shared limitation of the above-mentioned AILEd reviews is their focus on mapping research trends without providing an in-depth analysis of each included study. Additionally, most reviews did not cover publications in the past two years, a period which witnessed game-changing advancements in educational AI technologies due to the release of the Transformer algorithm and GenAI applications like ChatGPT. This omission would present a limited view of existing studies. Moreover, quality assessment of included studies was almost absent in existing reviews, causing problems in drawing firm conclusions. As Yang and Kyun (2022) acknowledged, the lack of rigorous quality appraisal of individual studies may introduce potential bias, diminishing the reliability and validity of the review findings.

In response to these research gaps, the current review aims to synthesise studies published since 1956 and include the latest publications on GenAI technologies. By collecting data from a broader timeframe and encompassing the recent breakthroughs, this review aims to capture an exhaustive overview of extant AILEd research. Additionally, a quality appraisal will be conducted to assess the trustworthiness of the collected evidence. Given the extensive empirical studies on the influence of AI applications in language learning (cf. Section 2.3) and the potential for enhanced effects with recent advancements in emerging AI technologies, this review delves into studies on AI-assisted language learning effectiveness. Specifically, the following research questions (RQs) will be addressed:

RQ1: What is the extent of research that has been conducted to investigate the effectiveness of AI-assisted tools in L2/FL learning?

RQ2[1]: What is the extent of the evidence of the effectiveness of AI-assisted L2/FL learning in higher education contexts from studies identified in RQ1?

2a. What are the effects of AI-assisted learning tools on language learners' learning outcomes?

2b. What are the effects of AI-assisted learning tools on language learners' learning perceptions?

[1] Of note, the scope of RQ2 has been narrowed for the feasibility of data synthesis within the specified time. Detailed reasons for this decision will be clarified in the section below (cf. Section 4.1).

Chapter 3 Method

This chapter first explains the reasons for conducting a systematic review to address the research questions. It then outlines the role of a protocol and describes planned methods of the review by presenting relevant sections from the protocol (see Appendix 1).

3.1 Why a systematic review was chosen?

A systematic review is a secondary analysis and rigorous synthesis of existing primary research on a particular topic, which aims to reveal the current state of knowledge and inform future research (Dickson et al., 2017). It differs from other types of reviews in several aspects: firstly, a systematic review attempts to be as exhaustive as possible by identifying all studies relevant to the topic, thereby minimising the selection bias in traditional methods of literature review, such as narrative reviews; secondly, a systematic review addresses reviewer bias by being transparent about the literature retrieval, inclusion, appraisal, and synthesis processes, which enhances the scientific rigour and reproducibility of the review; finally, a systematic review involves quality assessment of the included studies, which contributes to the reliability of review findings by including only methodologically sound studies (Macro, 2019). Altogether, the comprehensiveness, low risk of bias, and trustworthiness of a systematic review make it a robust methodological approach for this research.

In conducting the current systematic review, the author referred to Boland et al.'s (2017) student guidebook on systematic reviews and followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) guidelines to report items that should be included (Moher et al., 2009). Since this review is in the field of education, the author also consulted the International Database of Education Systematic Reviews (IDESR), a digital repository of published systematic reviews and pre-published review protocols in education, to locate all available reviews pertaining to the research topic and avoid duplication of efforts (Chalmers et al., 2023).

Before initiating the formal review process, a protocol was generated following the PRISMA-P statement (Moher et al., 2015) to guide how this systematic review would be conducted. The next section will outline how the review protocol was developed and registered.

3.2 Protocol creation and registration

A protocol pre-specifies how a systematic review is intended to be conducted, including the rationale, objectives, and *a priori* methods of the review (Moher et al., 2015). Generating and pre-registering a protocol is important for the following reasons: firstly, a protocol enhances the transparency of the

review process and ensures consistency among reviewers; secondly, it allows readers to identify modifications to the review method and detect review bias by comparing the planned content with the reported content; and lastly, it helps avoid duplication of reviews (Stewart et al., 2012).

For this review, a protocol was created and prospectively registered on IDESR (https://idesr.org/article/IDESR000120). The protocol was followed precisely as outlined unless otherwise specified in the text below. For example, RQ2 was revised from examining all phases of education to focusing solely on tertiary education, due to the larger-than-expected number of eligible studies. Section 3.10 and Section 4.1 will elaborate on the reasons for this revision.

The following sections delineate the review method and analysis procedures, including eligibility criteria, information sources, search strategy, data management, data screening and selection process, data extraction, risk of bias assessment, data synthesis, and confidence in cumulative evidence.

3.3 Eligibility criteria

The eligibility criteria were formulated based on the PICOSS elements (Cherry & Dickson, 2017). PICOSS stands for population, intervention, comparator, outcomes, study design, and setting. Due to the intricacy of research on the topic of interest, additional criteria elements, including completeness of reference, time of publication, language of publication, and type of publication, were considered in this review. Table 1 specifies the inclusion and exclusion criteria along with the rationale for each criterion.

Table 1. Eligibility criteria

Criteria	Inclusion	Exclusion	Rationale
Reference	Studies with complete bibliographic information	Studies with incomplete bibliographic information	The researcher needs a complete reference to retrieve the studies that are reviewed.
Time period	Studies published in or after 1956	Studies published before 1956	The concept of artificial intelligence was originally proposed in 1956 by John McCarthy.

Language of publication	Publications in English and Chinese	Publications in languages other than English and Chinese	The researcher is fully aware of the language bias caused by only including publications in certain languages. However, the researcher could only read English and Chinese.
Type of publication	Peer-reviewed literature and doctoral dissertations	Other publications that are not peer-reviewed (i.e., grey literature)	Peer-reviewed publications have been revised to achieve the best possible validity and accuracy before publication. Besides, Macaro (2019) suggested including doctoral theses because they have undergone rigorous scrutiny by examiners.
Participants	Studies on typically developing second and foreign language (L2/FL) learners of any age, gender, first language (L1) background, and L2/FL proficiency level	1) Studies on non-typically developing learners, such as those with Developmental Language Disorder; 2) Studies on typically developing L1 learners	Findings on non-typically developing learners may not be generalisable to the larger population. Additionally, preliminary searches have yielded few studies on using AI-assisted tools for L1 acquisition; thus, this review focuses on the effect of AI-powered applications on L2/FL learning.
Intervention	Studies that used all types of AI-supported learning tools (e.g., chatbots, intelligent tutoring systems, language learning applications equipped with AI technologies, etc.) to assist L2/FL learning	1) Studies that adopted AI-based learning tools to develop learners' L1 skills, such as literacy, and other skills unrelated to language acquisition; 2) Studies that used AI-based tools as an	The review does not focus on the effect of AI-assisted learning tools on L1 learners' learning achievements. Moreover, the review does not focus on language teachers' use or their perceptions of these applications.

		intervention to assist language teaching activities (e.g., assessment, feedback provision) or classroom management	
Outcomes	Studies reporting empirical data on the effectiveness of AI-assisted applications in L2/FL learning (e.g., learning outcomes measured by standardised tests, learners' attitudes collected by questionnaires and interviews)	Studies that did not report empirical data collected from L2/FL learners	The comparison of substantive learning outcomes (e.g., test scores, questionnaire data) could address the research question about the effectiveness of AI-assisted tools in second and foreign language learning
Study design	1) Randomized controlled trial (RCT) and non-RCT in which the experimental group(s) used AI to assist L2/FL language learning, while the control group(s) did not use AI; 2) Cohort study in which one group of participants adopting	1) Studies on the AI algorithms of learning applications that do not involve human participants; 2) Systematic reviews and meta-analyses of AI in L2/FL education	Experimental and quasi-experimental studies are most reliable for answering research questions on the effectiveness of AI-assisted language learning tools. Survey studies were also included because they were found relevant in preliminary scoping searches related to learning perceptions, although they may not add to findings on the effectiveness of these applications.

	AI-assisted language learning was pre-and-post tested and (/or) interviewed; 3) Cohort study in which one group of participants adopting AI-assisted language learning was only post-tested and (/or) interviewed; 4) Cross-sectional survey study in which data were collected at one point in time; 5) Mixed methods studies that combined the abovementioned study designs		
Setting	1) Studies in laboratory or classroom settings where L2/FL learners used AI-assisted tools; 2) Studies in which L2/FL learners used AI-assisted tools for self-study in various out-of-class settings	Studies in laboratory or classroom settings where AI-assisted tools were used to address teachers' needs	Laboratory and in- and out-of-classroom settings are included because the reviewed publications should be as comprehensive as possible.

3.4 Information sources

The reviewer conducted literature searches in nine databases covering the areas of education, linguistics, psychology and multidiscipline (see Table 2). These databases were identified after consultations with an experienced librarian at the Bodleian Education Library. Following Macro's (2019) suggestion, dissertation databases were also included, since doctoral theses mainly received rigorous evaluation from examiners. All databases were accessed electronically through Search Oxford Libraries Online (SOLO), the online catalogue for library collections at the University of Oxford.

In addition to searching electronic databases, the researcher followed Dundar and Fleeman's (2017) guidance to conduct forward and backward citation chaining to complement the search results. Once identifying a paper that met all inclusion criteria, the researcher conducted a forward search for eligible studies among subsequent publications that cited the paper. The researcher also examined the paper's reference list to look for relevant studies in backward searching (i.e. 'snowballing'). The researcher commenced literature searches on 25 April 2024 and updated the search results every two weeks. Considering the amount of time to complete the remaining stages of this project, the search was finalised on 31 May 2024.

Table 2. A list of consulted databases

Category	**Database**
Education	British Education Index
	Education Collection (incl. ERIC)
Linguistics	Linguistics Collection (including LLBA)
Psychology	PsycINFO
Multidiscipline	SCOPUS
	Web of Science Core Collection
Dissertation	ProQuest Dissertations and Theses
	Oxford University Research Archive (ORA)
	China Academic Journals (CNKI) - Chinese

3.5 Search strategy

Search items were generated based on previous systematic reviews on AI in education (e.g., Liang et al., 2023; Zheng et al., 2021). At the initial stage, the following keywords were identified for pilot searches: (1) words related to AI: "artificial intelligence" OR AI OR "artificial neural network" OR "machine intelligence" OR "machine learning" OR "deep learning" OR robotic OR "intelligent support" OR "intelligent virtual reality" OR chatbot OR "chat bot" OR "conversational agent" OR "automated tutor" OR "personal tutor" OR "intelligent agent" OR "intelligent system" OR "intelligent tutor" OR "natural language processing"; (2) words related to language learning: language OR "second language" OR L2 OR "foreign language" OR pronunciation OR vocabulary OR grammar OR listening OR speaking OR reading OR writing; (3) words related to effect: outcome* OR achieve* OR perform* OR effect* OR impact*. The asterisk was used as a truncation symbol to search for free-text words with the same root but alternative endings. For example, a keyword search for *achieve** would retrieve search results containing *achieve*, *achieves*, *achieved*, *achievement*, and *achievements*.

Following pilot searches with the author's supervisor and the librarian, two modifications were made to the search terms. Firstly, given the increasing prominence of generative AI and large language models, the supervisor suggested adding "generative artificial intelligence" OR "generative AI" OR GenAI OR "large language model" OR LLM OR ChatGPT OR GPT-3.5 OR GPT-4 to the first cluster of AI-related search terms. Secondly, to ensure a manageable amount of search output, "language" in the second group of search items was refined to "language learn*" OR "language education" according to the librarian's suggestion, because the term "language" easily yielded extraneous search results.

To keep the number of search results manageable and ensure data screening before the deadline, AI-related terms were assigned to the search frame of TITLE, while language learning and effect-related terms were assigned to the search frame of FULL TEXT. The author and the librarian jointly decided on these search frames after more than five search trials, which effectively reduced the number of irrelevant studies. Boolean operators "AND" and "OR" were used to connect the search items. Sample Boolean search strings in an English database and a Chinese database are presented in Appendix 2. Due to the interdisciplinary nature of the research topic, the Chinese search string was cross-checked by two Chinese-English bilingual speakers, one studying applied linguistics and the other studying computer science at the University of Oxford.

3.6 Data management

Three tools were used to manage the search data, including Rayyan, EndNote, and Microsoft Excel. Firstly, Rayyan was employed for efficient screening against the inclusion and exclusion criteria. Secondly, EndNote was used for full-text screening, as it allowed for text annotation and citation management. Lastly, Microsoft Excel was used to extract data and maintain records throughout the review process. The next section details how each tool was used to manage the retrieved data.

3.7 Data screening and selection process

3.7.1 Title and abstract screening

After conducting literature searches in relevant databases and through citation chaining, all search results (n = 3547) were exported as RIS files and uploaded to Rayyan (Ouzzani et al., 2016), a web tool for data screening and management in systematic reviews. The reviewer first performed duplicate detection on Rayyan to remove duplicate items (n = 675) and then examined the titles and abstracts of the remaining articles (n = 2872) to determine whether to include them or not according to predefined criteria (cf. Table 1). Articles that met all inclusion criteria were marked as "include". Articles that were yet to be decided on inclusion or exclusion were also tagged as "include". Articles that violated any inclusion criterion were marked as "exclude" (n = 2670), with reasons for their exclusion specified, including irrelevant topics (i.e., articles unrelated to language education, n = 1710), wrong intervention (i.e., articles that involved using AI-assisted tools for the acquisition of skills other than language or for language teaching, n = 506), wrong study design (i.e., articles focusing on AI mechanisms, systematic reviews, or meta-analyses, n = 245), wrong outcomes (i.e., articles lacking empirical data, n = 113), wrong population (i.e., articles involving language teachers, n = 79), background articles (n = 16), and articles not written in English or Chinese (n = 1).

3.7.2 Full-text screening

Following the initial screening, the articles labelled as "included" were retrieved for their full texts through SOLO and academic platforms such as ResearchGate. If the full texts were unavailable in these repositories, the reviewer contacted the authors to obtain their publications. Despite these efforts, 6 book chapters were unable to be retrieved due to Oxford's library copyright regulations that only one chapter can be scanned from the requested book. All available full texts (n = 196) were then imported to Endnote, a reference management tool, for screening. A Microsoft Excel spreadsheet was

used to record the decisions for inclusion or exclusion as well as the reasons for these decisions.

3.7.3 Quality assessment

To assess the reliability and validity of eligibility criteria, a second reviewer studying applied linguistics at the University of Oxford was recruited to participate in both phases of screening. Before screening, the second reviewer read through the review protocol to understand the review objectives and eligibility criteria. According to the protocol, the second reviewer was to screen 10% of the total search results in both screenings. However, due to the unexpectedly large amounts of retrieved data, the second reviewer only screened 5% of total articles in the first-round screening (n = 143) and 10% in the second-round screening (n = 20).

To minimize the risk of bias, the “Blind On” mode in Rayyan was used to ensure that the first reviewer’s decisions were invisible to the second reviewer. For the title and abstract screening, the inter-rater agreement was 98.6%, yielding a Cohen’s Kappa of 0.915. For the full-text screening, the inter-rater agreement was 95%, yielding a Cohen's Kappa of 0.857, exceeding the threshold of 0.7 for Cohen’s Kappa in systematic reviews in education (Frey, 2018). After comparing the decisions of both reviewers, any conflicts were discussed until a consensus was reached.

Following two rounds of screening, 105 articles were included in this review. To demonstrate the screening and selection process, a PRISMA flow diagram is illustrated in Figure 2.

Figure 2. Flow diagram of the data screening and selection process (adapted from Page et al., 2020)

3.8 Data extraction

A data extraction form (see Appendix 3) was created on Microsoft Excel to record relevant data that might answer the research questions, such as participants, interventions, and outcomes. Following Fleeman and Dundar's (2017) guide, the author piloted the form among three included studies with her supervisor to validate its effectiveness in data extraction. While the protocol predetermined that the second reviewer would extract 10% of the included studies, this procedure was not followed due to time constraints.

3.9 Risk of bias assessment

Quality assessment of individual studies was conducted after data extraction to reduce reporting bias. The 2018 version of the Mixed Methods Appraisal Tool (MMAT; Hong et al., 2018) was used to assess the methodological quality of the included studies. The quality assessment tool was chosen because it is suitable for appraising quantitative, qualitative, and mixed methods studies with different research designs, which facilitates standardised assessment across studies (Pace et al., 2012). Additionally, this appraisal tool has been used in previous systematic reviews in various topic areas (e.g., Schulz et al., 2023; Willis et al., 2019). Specifically, MMAT can be used with five categories of studies: (1) qualitative studies, (2) quantitative randomised controlled trials, (3) quantitative non-randomised studies, (4) quantitative descriptive studies, and (5) mixed methods studies. Five quality criteria are provided for assessing each of the five categories, and each criterion can be marked "yes", "no", or "can't tell". The MMAT and an explanation of each of its assessment categories are shown in Appendix 4. The second reviewer was asked to assess 10% of the included studies (n = 10).

3.10 Data synthesis

Due to the unexpectedly high number of included studies following two rounds of screening (n = 105), RQ2 was revised to focus on higher education (cf. Section 2.4). This decision was made after discussions with the author's supervisor. Further explanations for this choice will be provided in Section 4.1. Drawing from the data synthesis procedures in Macaro et al. (2012, 2018), a systematic mapping was conducted to answer RQ1, followed by an in-depth review of studies in tertiary education contexts (n = 71) to address RQ2.

For RQ1 concerning overall research characteristics, figures were generated using Excel and VOSviewer (Van Eck & Waltman, 2010), accompanied by qualitative descriptions. For RQ2, the synthesis followed Petticrew and Roberts' (2006) framework to present a narrative summary of the data. The three steps consisted of (1) grouping studies into two logical categories corresponding to the two sub-questions for RQ2, namely learning perceptions and learning outcomes; (2) analysing findings within each category; and (3) summarising findings across studies. According to the protocol, the included studies were checked for their homogeneity to decide whether it was appropriate to combine the studies in a meta-analysis. However, due to the heterogeneity of the included studies, meta-analysis was not conducted in this review.

3.11 Confidence in cumulative evidence

The overall quality of evidence (also referred to as “the certainty of evidence” or “the confidence in the effect estimates”) was evaluated with the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach (https://www.gradeworkinggroup.org). The reviewer considered five down-rating criteria: (1) risk of bias in individual study design, (2) inconsistency of results across studies, (3) indirectness of evidence (i.e., limitations in the generalisability of results), (4) imprecision (i.e., wide confidence intervals around estimates of effect), and (5) publication bias. Criteria for rating up may also be applied when there exist (1) large effect sizes, (2) a dose-response gradient, or (3) plausible residual confounding. After discussion with the second reviewer, the result would be categorised as very low, low, moderate, or high quality of evidence.

Chapter 4 Results

Following the two steps of data synthesis suggested by Macaro et al. (2012, 2018) (cf. Section 3.10), this section first summarises the overall characteristics of the 105 eligible studies and then provides an in-depth synthesis of 71 studies within higher education contexts.

4.1 Mapping of included study characteristics

To answer RQ1 on the extent of existing research on AI-assisted L2/FL learning, this subsection summarises all included studies (n = 105) from publication trends, participant characteristics, geographical distribution, adopted AI technologies, studied language skills, types of study design, and theoretical backgrounds. Appendix 5 lists detailed information for each study and its corresponding ID. The summary is followed by a risk of bias assessment of individual studies and keyword analysis.

4.1.1 Publication trends

Figure 3 presents the distribution of all included studies according to their year of publication. According to the line chart, the earliest eligible study was published in 1995, marking an initial rise in research interest in AI-assisted language learning. There was a lack of published works between 1996 and 2013, with sporadic research emerging from 2014 onwards. The area remained relatively under-investigated until 2019 when the number of publications began to steadily increase. Over the past four years, there has been a noticeable surge in the number of research outputs, indicating a burgeoning interest in this field. The decline in the number of publications in 2024 is because the literature search ended on 31 May 2024; thus, only studies from the first half of the year were included. The total number of research outputs is anticipated to grow in 2024, continuing the upward trend in recent years.

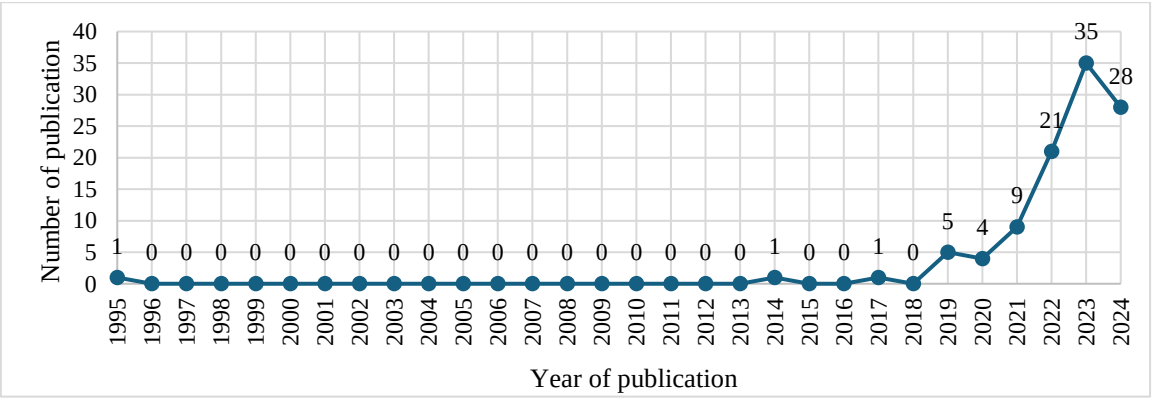


Figure 3. Publication trends

4.1.2 Participant characteristics

Participant characteristics are summarised in terms of sample size, gender distribution, educational levels, and language backgrounds.

According to Appendix 5, the number of participants in the included publications ranged from 2 (Study No. 62) to 1 million (Study No. 44). Appendix 5 also shows that 57 of the 105 studies provided gender information, with 2 studies exclusively involving male participants and 3 studies including only female participants. Of the remaining 52 studies, 15 were male-dominant, 35 were female-dominant, and 2 had an equal gender proportion.

Figure 4 shows the number of included studies across different educational phases. The majority of studies (71 studies, 68%) involved tertiary-level students, followed by elementary (14 studies, 13%) and secondary school students (12 studies, 11%). Additionally, 8 studies were conducted among participants with varying educational backgrounds.

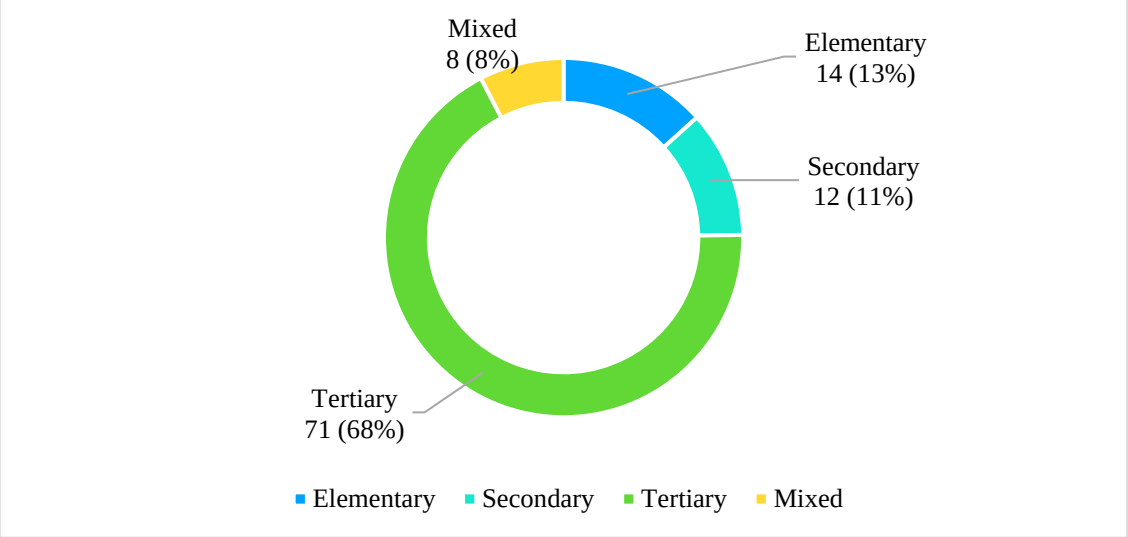


Figure 4. Participants' educational levels

Regarding the participants' language backgrounds, the three most common L1s among participants were Chinese (36 studies), Korean (11 studies), and Arabic (6 studies) (see Figure 5). Figure 5 also illustrates the frequency of publications investigating different target L2/FL. It can be found that the overwhelming majority of studies focused on English learning, amounting to 92 studies. Chinese and Japanese were the second and third most studied languages, with 4 and 2 studies contributing to each language. Other target languages included Arabic, French, German, Korean, Polish, Russian, and Thai. Participants' L2/FL proficiency was revealed in 78 studies, of which 5 involved beginners, 22 involved intermediate learners, 50 involved participants with mixed language proficiencies, and only one involved advanced learners (see Figure 6).

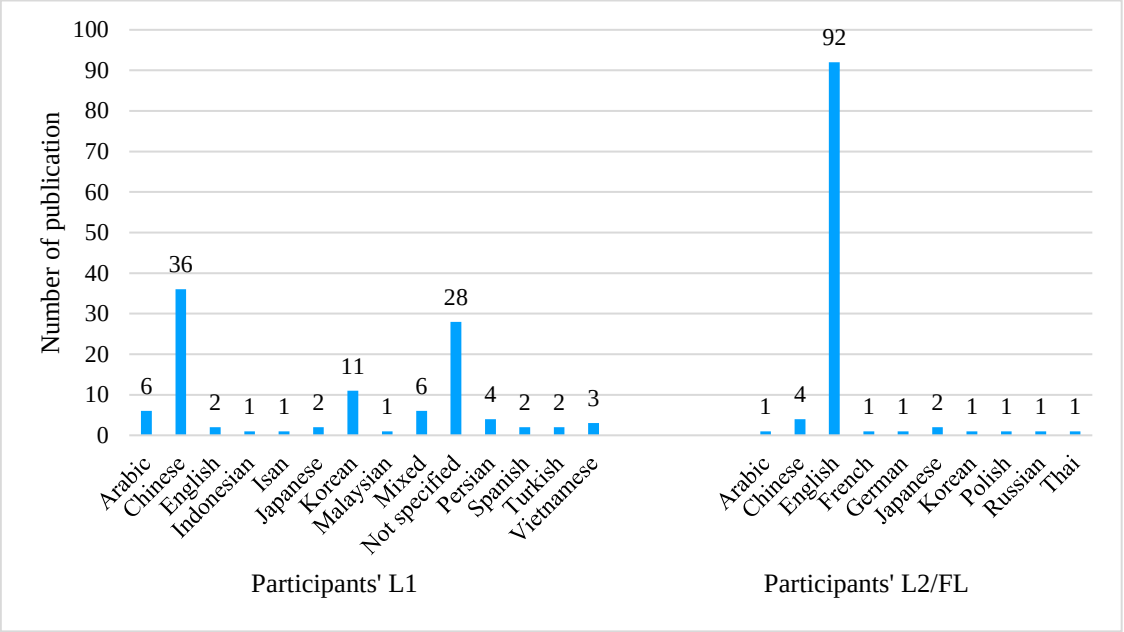


Figure 5. Participants' L1 and L2/FL

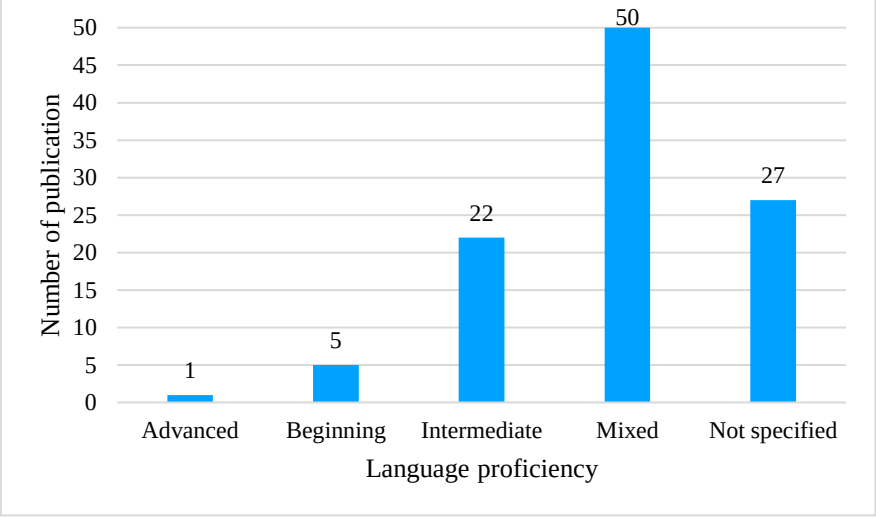


Figure 6. Participants' L2/FL proficiency

4.1.3 Geographical distribution

Figure 7 depicts that the included studies were conducted across 23 different countries. The largest share, totalling 40 publications, was carried out in China. Following this, Korea, Iran, and India contributed 12, 7, and 5 studies respectively. Other contributing countries included Algeria, Egypt, Ethiopia, Germany, Indonesia, Japan, Kazakhstan, Malaysia, Mexico, Norway, Russia, Saudi Arabia, Singapore, Spain, Thailand, Turkey, Ukraine, the United States, and Vietnam. Additionally, two studies did not specify their location.

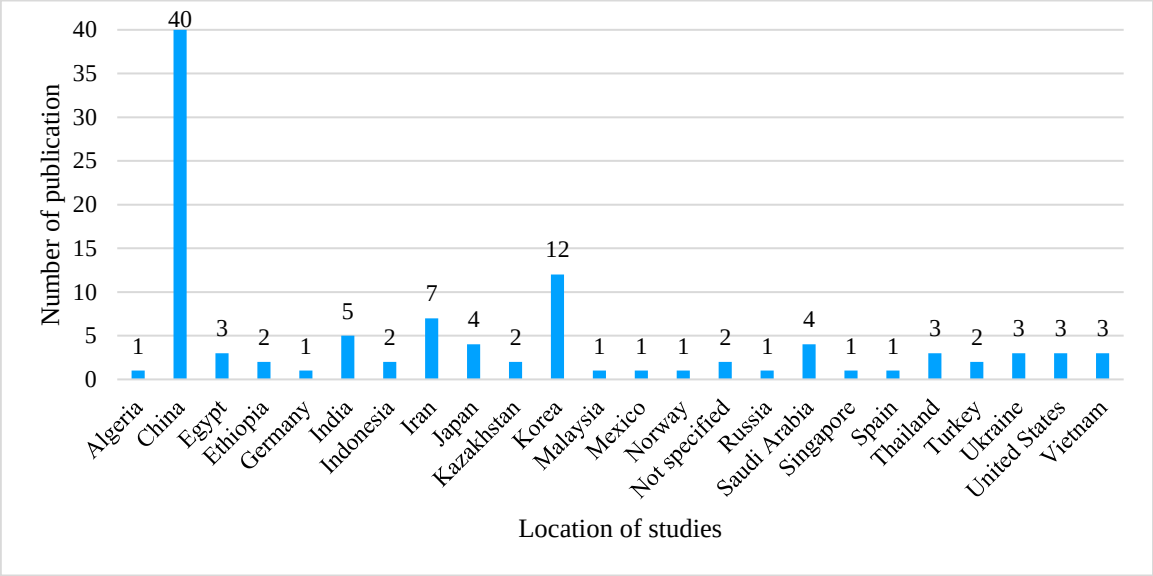


Figure 7. Location of study

4.1.4 Adopted AI technologies

The included publications investigated various AI technologies, which can be categorised into five types (see Figure 8). Specifically, the majority of studies (45 studies, 43%) examined chatbots, such as ChatGPT, Ellie, and Replika. This was followed by 28 studies (27%) on AI-assisted language learning applications or platforms (e.g., Duolingo and Speeko) and 16 studies (15%) on AI-assisted writing and evaluation tools (e.g., Grammarly). The remaining studies focused on intelligent tutoring systems (13 studies, 12%) and speech recognition tools (3 studies, 3%).

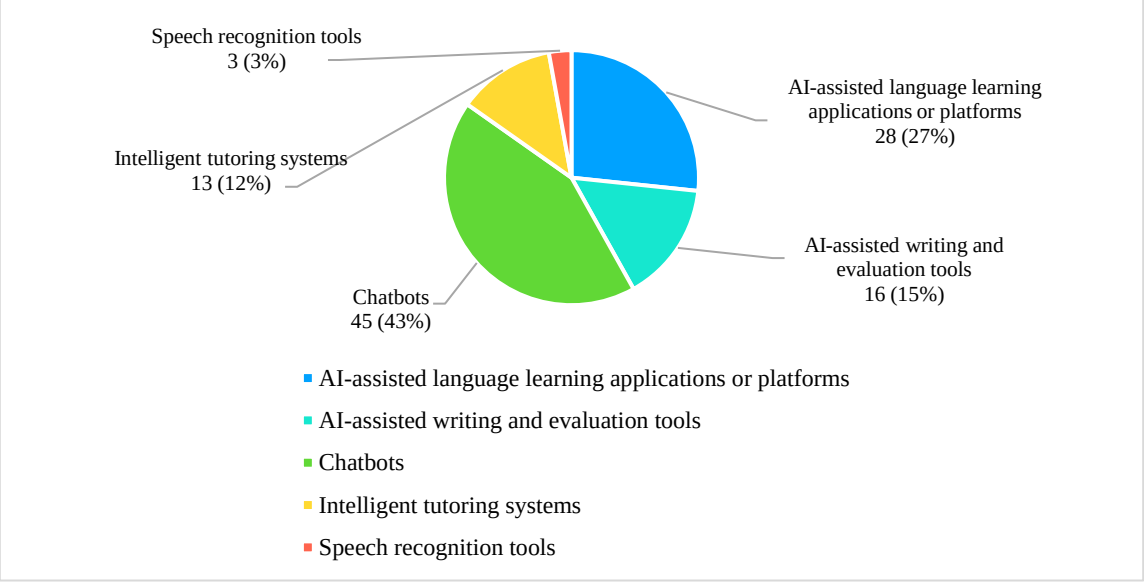


Figure 8. Types of AI technologies

4.1.5 Studied language skills

Figure 9 displays the number of publications targeting different language skills. 26 studies respectively investigated writing skills and integrated language skills respectively, followed by studies on speaking (24 studies), vocabulary (11 studies), reading (8 studies), grammar (4 studies), pronunciation (3 studies), and listening (3 studies).

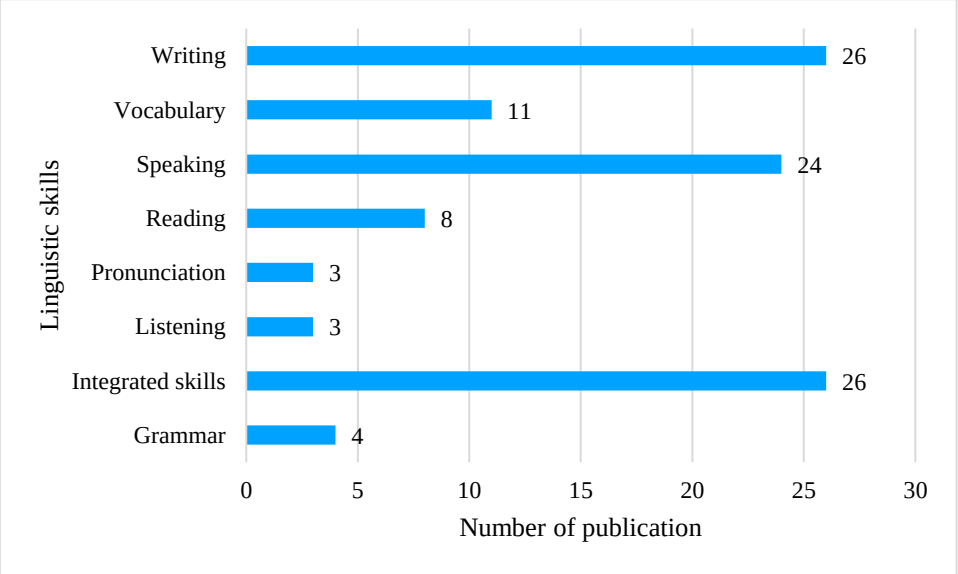


Figure 9. Target language skills

4.1.6 Types of study design

Figure 10 demonstrates that the study design of the included publications was relatively diverse, with more than half studies adopting a mixed-method approach (n = 60). The remaining studies consisted of 39 quantitative studies and 6 qualitative studies.

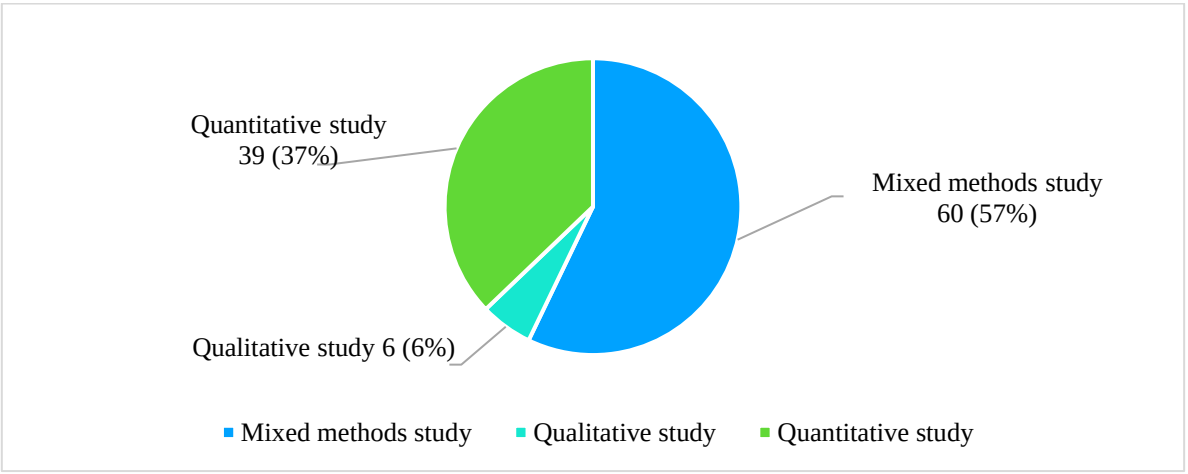


Figure 10. Types of study design

4.1.7 Theoretical backgrounds

Figure 11 depicts the theoretical frameworks underpinning the included studies. Notably, 69 studies did not specify their theoretical foundations, accounting for approximately 65% of the total. The most frequently used theory is the sociocultural theory (*n* = 11), followed by the technology acceptance model (*n* = 5) and the engagement theory (*n* = 3). Cognitive SLA theories were also prominent, including the Input Hypothesis (*n* = 2), the Interaction Hypothesis (*n* = 2), and the Noticing Hypothesis (*n* = 1).

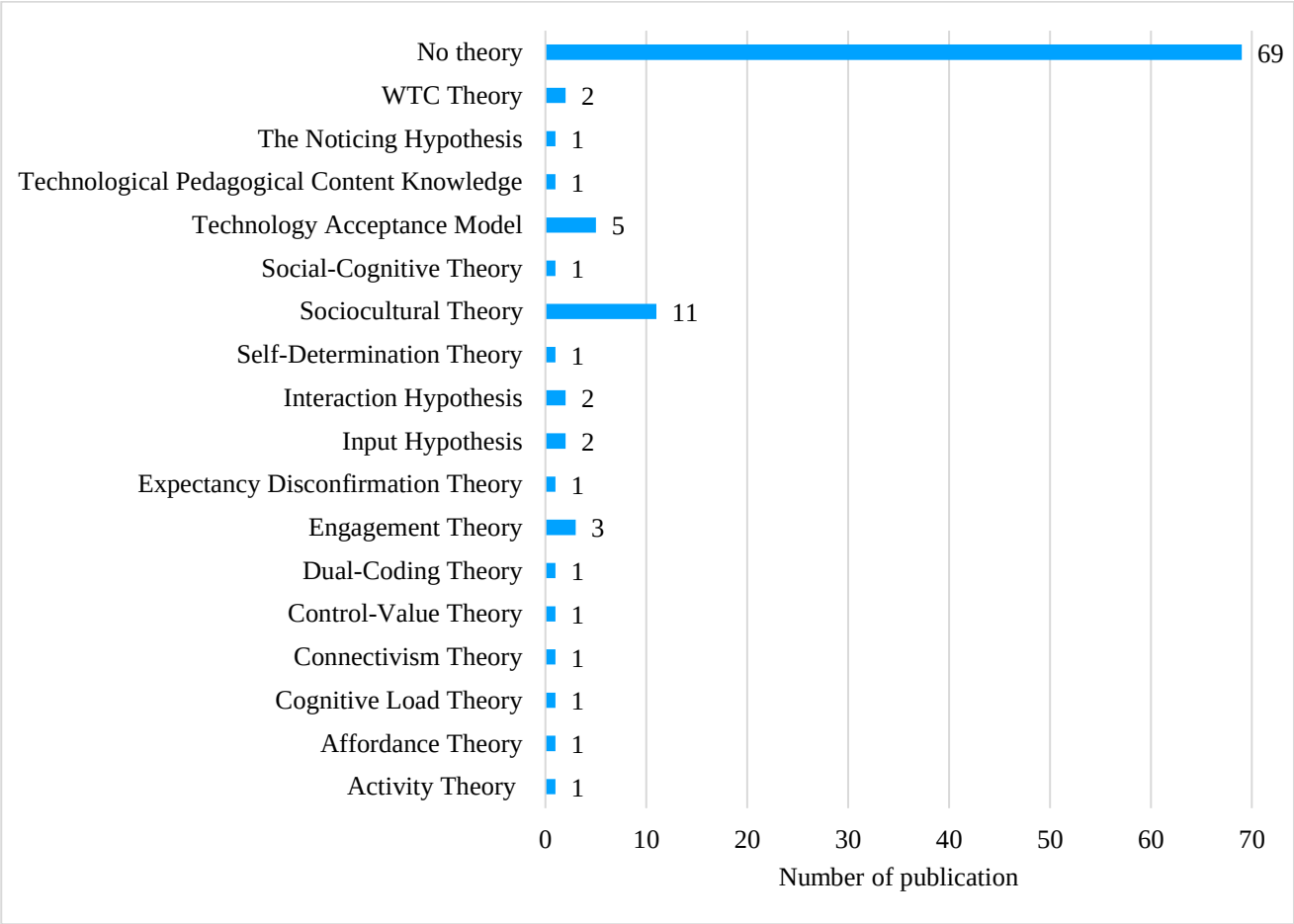


Figure 11. Theoretical groundings of included studies

4.1.8 Risk of bias assessment

Appraised by the MMAT, the risk of bias (RoB) for individual studies is presented in Appendix 6. Three studies received “high”, 78 “moderate”, and 24 “low” RoB ratings. The primary source of bias was confounding variables (*n* = 54), which affected the overall robustness of conclusions.

4.1.9 Keyword visualisation

Figure 12 is plotted in VOSviewer to show the top 15 keywords in the included studies. Keyword mapping echoes several study characteristics discussed above. Firstly, keywords such as *college students* (occurrences = 4) suggest a prominent research interest in the tertiary context. Secondly, *chatbots* (occurrences = 5) as a keyword reflects scholarly attention on this type of AI technology. Thirdly, the keyword *vocabulary* (occurrences = 9) implies the role of vocabulary as the building block for language learning and the foundation for all language skills.

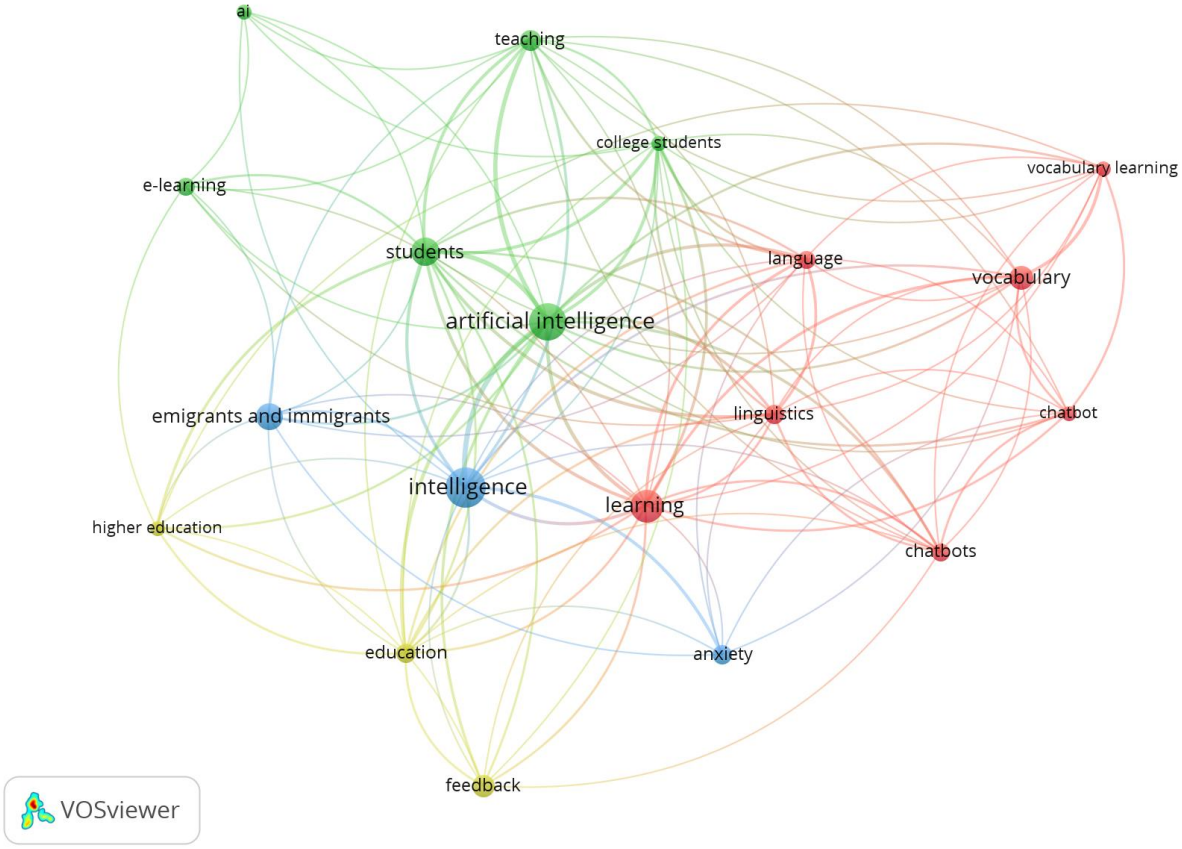


Figure 12. Keyword mapping

In summary, mapping the 105 included studies revealed the following characteristics:

1. The number of studies on AI-assisted language learning increased over the past three decades, particularly between 2020 and 2024.

2. There was a significant variability in sample size and gender distribution across studies, suggesting that the participants were recruited from a diverse pool. The participants were primarily tertiary-level students studying English as L2/FL. In some studies, the participants' L1 and L2/FL proficiencies were not explicitly reported.

3. Studies were predominantly conducted in Asian countries, such as China and Korea.
4. The most frequently investigated AI technologies were chatbots.
5. Integrated and writing skills attracted the researchers' most attention, while pronunciation and listening skills were less studied.
6. The largest number of studies adopted mixed methods, followed by quantitative studies and qualitative studies.
7. A large portion of studies did not delineate their theoretical backgrounds, revealing a need to base studies on theoretical frameworks.
8. Most studies were rated "moderate" RoB, suggesting that their findings should be interpreted with caution due to methodological flaws.
9. Keyword mapping reveals a research trend in using AI technologies in higher education contexts.

Following an overview of studies across all educational levels, the next section delves into research in tertiary education contexts to answer RQ2. The recent flourishing of AI-assisted language learning research resulted in a greater number of eligible studies than anticipated, which posed challenges to data synthesis before the deadline for this project. Following Macaro et al.'s (2012, 2018) approach to concentrating on specific educational phases, the author discussed with her supervisor and selected higher education as the focus. The decision was driven by the observation that most studies involved university or college language learners (cf. Section 4.1.2), which contributed to the feasibility of data synthesis within the given time constraints. The author acknowledged that narrowing the RQ scope at this stage deviated from the protocol and introduced selection bias. More critical reflections will be presented in Section 6.3.

4.2 In-depth analysis of studies in higher education

Considering the deadline for this project, the in-depth analysis only reported 71 studies in higher education settings, since they accounted for the largest proportion of all eligible studies.

To answer RQ2a on the extent of evidence of AI-assisted language learning outcomes, the following analysis groups relevant studies by the language skills they addressed. Substantive learning outcomes of three types of linguistic knowledge (i.e., vocabulary, grammar, and pronunciation) will be analysed first, followed by synthesis of learning outcomes of four language skills (i.e., listening, speaking, reading, and writing) and integrated language skills.

These studies mostly employed a pretest-postest design and measured tangible learning outcomes by comparing pretest scores with post-test scores. Some studies are controlled trials that compare an experimental group (EG) against a control group (CG), while others involved one cohort

compared before and after the intervention. Study results are labelled "positive", "negative", "neutral", or "mixed", with definitions for each category provided in Table 3. Extracted data from these studies are presented in Table 4 to Table 12 below.

Table 3. Definition of learning outcome categories

	Controlled trials	**Pre-post study**
Positive	Significantly higher post-test scores in the EG than CG	Significant increase in scores from pre- to post-intervention
Negative	Significantly lower post-test scores in the EG than CG	Significant decline in scores from pre- to post-intervention
Neutral	No significant differences in post-test scores between EG and CG	No significant differences in scores from pre- to post-intervention
Mixed	Significant improvements in some measures between EG and CG, but decline or no difference in others	Significant improvements in some measures from pre- to post-intervention, but decline or no difference in others

4.2.1 Evidence of learning outcomes from vocabulary studies

Four studies on vocabulary learning adopted a pretest-posttest control group design to report learning outcomes by comparing test scores between CGs and EGs. Table 4 displays the data extracted from these studies.

Two of the four studies investigated the effect of chatbots on vocabulary acquisition. Both studies yielded large positive effect sizes, suggesting that chatbots, with their capabilities to provide real-time interaction and tailored feedback, could enhance vocabulary learning. The first study (Qasem et al., 2022) involved 40 male tertiary-level students in Saudi Arabia in a 12-week randomised controlled trial (RCT). The participants were divided into a CG receiving traditional teacher-led instruction and an EG receiving the same instruction supplemented with chatbot assistance for checking vocabulary meaning and providing contextual examples. Results indicated that both groups improved their vocabulary scores from pre-test to post-test, with the EG showing significantly greater gains. The second study (H. Chen et al., 2020) examined a chatbot capable of providing timely feedback and repetitive exercises. The effectiveness of the chatbot in Chinese vocabulary learning was assessed through a four-week non-randomised comparison among 58 university students. The CG used the chatbot in one-on-many classroom settings, while the EG used the chatbot in one-on-one learning environments. Results revealed that both groups improved their post-test scores, with the EG achieving better outcomes.

Huang and Wang (2023) performed an RCT among 132 Taiwanese university students to evaluate the impact of an AI-based motion sensing system on French vocabulary learning. The intelligent system could detect gestures and provide immediate feedback, having the potential to facilitate knowledge acquisition through body movements. It was found that participants in both groups learning with or without the system achieved better scores in the post-test, with those using the system showing significantly more vocabulary gains and word recall.

The remaining study (Agnes & Srinivasan, 2024) explored the integration of ChatGPT-generated mnemonic keywords into *Anki*, a flashcard application equipped with spaced repetition to enable review at optimal intervals and facilitate long-term retention. Pre-test and post-test scores demonstrated that both mnemonic and non-mnemonic groups improved significantly in vocabulary retention, with the mnemonic group outperforming the other group.

Table 4. Extracted data from vocabulary studies

Reference	**CG**		**EG**		**Significance values and effect sizes**		**Result**
	Mean pre-test score	**Mean post-test score**	**Mean pre-test score**	**Mean post-test score**	**Pre/post-test**	**Between groups**	
Agnes & Srinivasan (2024)	24.33	38.46	24.4	41.93	CG: $p < .001$ EG: $p < .001$	$p < .01$	Positive
H. Chen et al. (2020)	71.63	88.03	49.82	91.33	CG: $d = 1.18$ EG: $d = 2.95$	$p = .000$	Positive
Huang & Wang (2023)	/	9.64	/	10.95	/	$p < .001$	Positive
Qasem et al. (2023)	9.35	11.7	9.05	14.85	/	$p = 0.006$ $\eta^2 = 0.426$	Positive

While all the above-reviewed studies established the effectiveness of AI-assisted tools in vocabulary learning, the generalisation of findings may be limited by divergence in outcome measures across studies. All four studies utilised self-developed vocabulary tests to measure vocabulary gains, but they did not provide sufficient detail on test design to assess their reliability and validity. Furthermore, although some studies (Agnes & Srinivasan, 2024; Huang & Wang, 2023)

generated statistically significant results, they did not report effect sizes, which failed to provide a measure of the magnitude of effect.

4.2.2 Evidence of learning outcomes from grammar studies

Two quantitative RCTs (see Table 5) reported substantive effectiveness of AI technologies in grammar acquisition. Nagata (1995) investigated a self-developed language instruction system for learning Japanese sentence constructions. The researcher compared the system that detected grammatical errors and provided explanations (EG) with traditional tutoring systems that solely identified errors (CG). Similarly, Nghi et al. (2019) examined the use of chatbots as supplementary tools in classroom instruction. The EG received English preposition instructions alongside interactive feedback from chatbots, while the CG only received classroom instructions. According to statistics in Table 5, both studies indicated substantial improvements in grammatical knowledge for the CG and EG, with the EG showing more significant gains, demonstrating the potential of AI-assisted language learning. The "low" RoB ratings for both studies reflected the trustworthiness of findings.

Similar to those vocabulary studies, the grammar studies did not report effect sizes. Additionally, further research is warranted to explore the effect of various AI technologies on learning different grammatical topics and to generalise the findings.

Table 5. Extracted data from grammar studies

Reference	**CG**		**EG**		**Significance values and effect sizes**		**Result**
	Mean pre-test score	**Mean post-test score**	**Mean pre-test score**	**Mean post-test score**	**Pre/post-test**	**Between groups**	
Nagata (1995)	52.9	67.5	53.8	80.1	/	$p = 0.02$	Positive
Nghi et al. (2019)	/	Unit 1 = 6.21 Unit 2 = 6.32 Unit 3 = 6.18 Unit 4 = 6.21 Unit 5 = 6.33 Unit 6 = 5.94 Unit 7 = 6.05	/	Unit 1 = 6.71 Unit 2 = 7.53 Unit 3 = 7.23 Unit 4 = 6.66 Unit 5 = 6.85 Unit 6 = 6.83 Unit 7 = 6.95	/	$p < .001$	Positive

4.2.3 Evidence of learning outcomes from pronunciation studies

One study (see Table 6) focused on pronunciation acquisition. El-Zeiny et al. (2023) conducted a pretest-posttest control group study among 49 tertiary-level students in Egypt to examine the effectiveness of an AI-driven storytelling tool. Both CG and EG received classroom instruction on Arabic emphatic consonant pronunciation over six weeks, with the EG assisted by additional practice and personalised feedback offered by the AI tool. Results showed that the EG achieved significantly better pronunciation skills test scores than the CG. However, the study involved a potential confounding factor, since the EG's better score may be partially attributed to the additional time on pronunciation practice rather than the AI-driven approach itself. The study was, therefore, rated "moderate" RoB. To address this problem, future studies should ensure equal instructional time for both groups. Future works are also expected to involve larger sample sizes, assess more speech sounds, and extend intervention periods to provide more insights into the effectiveness of ASR technology.

Table 6. Extracted data from pronunciation studies

Reference	CG		EG		Significance values and effect sizes		Result
	Mean pre-test score	**Mean post-test score**	**Mean pre-test score**	**Mean post-test score**	**Pre/post-test**	**Between groups**	
El-Zeiny et al. (2023)	10.2	11.38	9.8	16.35	CG: $p = 0.00$ EG: $p = 0.00$	Pre-test: $p > 0.05$ Post-test: $p < 0.05$, $\eta^2 = 0.5615$	Positive

4.2.4 Evidence of learning outcomes from listening studies

Three studies examined the role of AI-assisted tools in developing L2/FL listening skills, reporting mixed findings (see Table 7). Liao (2023) studied the effect of an intelligent teaching platform on 93 Chinese college students' listening skills. The participants received pre-course vocabulary preview, in-course intensive listening exercises, and post-course automated feedback on the platform for 18 weeks. Results showed that the students, regardless of language proficiencies, significantly improved their listening proficiency from the pre-test to the post-test. However, due to the lack of a CG, it should be cautioned that the effectiveness might simply be attributed to natural learning progressions after continuous instruction rather than the AI intervention.

Table 7. Extracted data from listening studies

Reference	CG		EG		Significance values and effect sizes		Result
	Mean pre-test score	**Mean post-test score**	**Mean pre-test score**	**Mean post-test score**	**Pre/post-test**	**Between groups**	
Liao (2023)	/	/	A-level = 65.13 B-level = 48.64	A-level = 79.89 B-level = 64.51	A-level: $p < 0.01$ B-level: $p < 0.01$	/	Positive
Li & Peng (2021)	84.552	71.241	85.167	71.950	/	Pre-test: $p = 0.709$ Post-test: $p = 0.780$	Negative
Vu et al. (2022)	Mean Difference (Posttest – Pretest) = 1.04		Mean Difference (Posttest – Pretest) = 3.3		CG: $p = 0.171$ EG: $p = .000$	Post-test: $p < .05$	Positive

Vu et al. (2022) investigated the impact of an AI-powered language learning application featuring movie caption transcription exercises on EFL listening comprehension in a 12-week RCT among 53 Vietnamese university students. The EG used the application for listening transcription tasks, corrective feedback, and follow-up personalised practices, while the CG completed traditional listening course book exercises. Results indicated that both groups enhanced listening comprehension in the post-test, but only the EG showed significant gains. However, the study was rated "high" RoB, due to insufficient consideration for the homogeneity of participants' pre-intervention English levels and lack of assessor blinding. Reporting effect sizes can be helpful to assess whether the statistically significant results were practically meaningful or only artefacts of flawed study designs.

By contrast, the study by Li and Peng (2021) failed to demonstrate the efficacy of AI interventions. The researchers employed a non-randomised comparison design, dividing 59 Chinese freshmen into an EG receiving personalised instruction on an AI-assisted English listening platform and a CG receiving traditional teacher-led face-to-face instruction. The study revealed lower post-test scores in both groups, with no statistically significant differences between them.

Overall, the effectiveness of AI-assisted tools in developing listening skills remained inconclusive, which might be due to different outcome measures across studies. Of the three studies, only Vu et al. (2022) utilised a standardised test — the Cambridge First Certificate listening tests — to assess general English proficiency. The other two studies (Li & Peng, 2021; Liao, 2023) did not specify their test design and appeared to adopt self-developed tests. This raised concerns about the reliability and validity of these tests for assessing the participants' language performance. The trustworthiness of findings was also limited by flaws in study designs. Another problem is the

omission of reporting effect sizes, which has been discussed in the previous sections for vocabulary and grammar studies.

4.2.5 Evidence of learning outcomes from speaking studies

13 speaking studies (see Table 8) reported findings on the learning effectiveness of AI-assisted language learning tools.

9 of the 13 studies focused on chatbots designed for speaking practice, which can be categorised into three strands. The first strand (Çakmak, 2022; Duong & Suppasetseree, 2024; El Shazly, 2021; Kim et al., 2021a) adopted a one-group pretest-posttest design in different EFL settings, including Turkey, Egypt, Vietnam, and Korea. These studies consistently established the effectiveness of chatbots in enhancing speaking performance after the intervention. For example, Kim et al. (2021a) assessed 49 Korean university students with different language proficiencies. Results showed that both low and intermediate-level groups improved significantly in pronunciation, intonation, and stress from the pre-test to the post-test, suggesting the effectiveness of the AI tool for users at varying proficiency levels. However, the absence of CG posed challenges in controlling confounders such as the natural progression of language learning, leading to "moderate" RoB in these studies.

The second strand consists of four studies (Fathi et al., 2024; M.-H. Hsu et al., 2023; Kemelbekova et al., 2024; Kim et al., 2021b) comparing chatbot-assisted groups with control groups learning without AI assistance. Employing a pretest-posttest design, all these studies indicated improvements in speaking skills from pretests to post-tests within groups, with the EG exhibiting more significant gains. While other studies involved two groups of participants, Kim et al.'s (2021b) investigation included 110 participants separated into three groups for English speaking practice — a face-to-face group where students interacted with their peers, an AI text-chatting group where students had written interaction with chatbots (Replika, Andy, and Google Assistant), and an AI voice-chatting group where students had verbal interaction with chatbots. It was found that both AI groups significantly improved their speaking performance in all assigned tasks (picture description, question responding, and opinion expressing), but the face-to-face group only showed significant improvements in the last two tasks. The comparison among groups revealed that the AI voice-chatting group significantly outperformed the other two groups in the task of expressing opinions.

The third strand features one study (Lin & Mubarok, 2021), in which participants utilised chatbots with different functionalities. The study investigated the effect of integrating mind maps into AI chatbots in a five-week intervention, with an EG using the mind map-guided AI chatbot and a CG using the traditional chatbot. It was identified that both groups exhibited improved speaking test results, with the EG benefiting significantly more, reflecting the positive impact of mind maps.

The remaining studies on language learning applications all included CGs and EGs but differed in their interventions. Three of the four studies involved a CG receiving traditional teacher-led instruction and an EG receiving AI-assisted learning on Duolingo (Qiao & Zhao, 2023), Speeko (Shafiee Rad, 2024), or FIF oral English Platform (Wu, 2022). All studies showed notably greater improvements in the EG after 10-week to 13-week intervention. Of note, while Wu (2022) stated "a significant increase in the students' average scores" (p. 73), the study did not report significance values or effect sizes, which undermined the reliability of the conclusion. Another study (Zou et al., 2023) targeted the influence of interactive activities, such as peer evaluation, in chatbot-assisted speaking practice. Both groups using chatbots with or without interactive activities attained better speaking scores following 5-week practice, with participants engaging in interactive activities scoring significantly higher.

Collectively, speaking studies highlighted the potential usefulness of chatbots to improve speaking skills. However, caution should be applied while interpreting the findings, since the studies adopted various speaking measures. While several studies focused on sub-measures of speaking, such as fluency, accuracy, and pronunciation (Duong & Suppasetseree, 2024; Fathi et al., 2024; Kim et al., 2021a; Qiao & Zhao, 2023), other studies only reported an overall score without explaining the marking descriptors. Future investigations may adopt complexity, accuracy, and fluency as indicators of speaking performance. These measures, which have been widely used for measuring language learning outcomes in applied linguistics (Ellis & Barkhuizen, 2005; Skehan, 1998), can enhance the comparability of findings across studies.

Table 8. Extracted data from speaking studies

Reference	**CG**		**EG**		**Significance values and effect sizes**	**Result**
	Mean pre-test score	**Mean post-test score**	**Mean pre-test score**	**Mean post-test score**		
Çakmak (2022)	/	/	69.7	74.3	Pre/post-test: p = .000	Positive
Duong & Suppasetseree (2024)	/	/	Mean Difference (Post-Pre) = .8167		Pre/post-test: p = .000	Positive
El Shazly (2021)	/	/	5.4	8.0	Pre/post-test: p = .001	Positive
Fathi et al. (2024)	5.12	6.67	4.98	5.89	Between groups: η_p^2 = .46	Positive

M.-H. Hsu et al. (2023)	75.42	73.21	76.79	141.25	Between groups: Pretest: p = .742 Post-test: p = .000	Positive
Kemelbekova et al. (2024)	15.96	15.56	15.69	19.69	Between groups: Post-test: p < 0.01	Positive
Kim et al. (2021a)	/	/	Mean Difference (Post-Pre): Low-level = 0.36 Intermediate = 0.25		Pre/post-test: Low-level: p < 0.01 Intermediate: p < 0.01	Positive
Kim et al. (2021b)	Task 1 = 1.51 Task 2 = 4.39 Task 3 = 2.67	T1 = 1.59 T2 = 4.97 T3 = 2.91	AI Text-Chatting: T1 = 1.30 T2 = 4.47 T3 = 2.69 AI Voice-Chatting: T1 = 1.32 T2 = 4.55 T3 = 2.91	AI Text-Chatting: T1= 1.49 T2 = 5.16 T3 = 2.94 AI Voice-Chatting: T1 = 1.69 T2 = 5.36 T3 = 3.53	Pre/post-test: CG: Task 1: p = .21 Task 2: p = .00 Task 3: p = .00 EG1: Task 1: p = .01; Task 2: p = .00; Task 3: p = .00 EG2: Task 1: p = .00; Task 2: p = .00; Task 3: p = .00 Among groups: Task 1: p = .21 Task 2: p = .49 Task 3: p = .00	Positive
Lin & Mubarok (2021)	6.75	6.90	6.80	8.16	Between groups: η^2 = 0.57	Positive

Qiao & Zhao (2023)	F = 5.24 V = 4.73 A = 5.32 P = 4.63	F = 5.52 V = 5.09 A = 4.97 P = 5.13	F = 5.12 V = 4.85 A = 5.26 P = 4.77	F = 5.94 V = 5.78 A = 6.32 P = 5.91	Between groups: F: p = 0.006 V: p = 0.034 A: p = 0.013 P: p = 0.025	Positive
Rad (2024)	5.26	6.12	5.12	8.25	Between groups: η_p^2 = 0.177	Positive
Wu (2022)	7.46	7.59	7.32	7.99	/	Positive
Zou et al. (2023)	5.24	5.84	5.42	6.58	Between groups: η^2 = 0.04	Positive

Note: F = Fluency, V = Vocabulary, A = Accuracy, P = Pronunciation

4.2.6 Evidence of learning outcomes from reading studies

One reading study (see Table 9) reported learning outcomes. Kim and Cha (2023) used a pretest-posttest research design to examine the role of AI translators in reading comprehension among 113 Korean university students over one semester. The participants were divided into three groups: one receiving teacher-led instructions to analyse reading texts, another using AI translators along with teacher-led instructions, and the third receiving revised AI translations alongside teacher-led instructions. It was found that all groups significantly improved their reading comprehension scores; yet, there were no significant differences among groups. The result suggested that the observed improvement may be attributed to factors other than AI interventions, such as the natural progression of learning. Calculating effect sizes could provide a more comprehensive understanding of the magnitude of the effect, informing readers of the most effective intervention for language development.

Table 9. Extracted data from reading studies

Reference	**CG**		**EG**		**Significance values and effect sizes**		**Result**
	Mean pre-test score	**Mean post-test score**	**Mean pre-test score**	**Mean post-test score**	**Pre/post-test**	**Between groups**	
Kim & Cha (2023)	55.6	66.89	EG1 = 62.38 EG2 = 61.38	EG1 = 69.13 EG2 = 70.25	CG: p < 0.01 EG1: p < 0.01 EG2: p < 0.01	Pre-test: p = .08 Post-test: p = .479	Neutral

4.2.7 Evidence of learning outcomes from writing studies

19 studies investigated L2/FL writing learning outcomes, of which 7 focused on chatbots and 12 on AI-assisted writing and evaluation tools, all reporting positive outcomes. Outcome data from studies on the two types of AI technologies is presented in Table 10 and Table 11.

Table 10. Extracted data from writing studies on chatbots

Reference	CG		EG		Significance values and effect sizes		Result
	Mean pre-test score	Mean post-test score	Mean pre-test score	Mean post-test score	Pre/post-test	Between groups	
Boudouaia et al. (2024)	/	13.41	14.70	18.78	EG: d = 0.77	p = .00	Positive
Ghafouri et al. (2024)	/	3.34	/	4.40	/	p < .001 η^2 = .50	Positive
Mahapatra (2024)	/	16.371	/	19.216	d= -3.300	p = 3.297e-7	Positive
Song & Song (2023)	37.31	45.18	39.26	59.12	/	p < 0.001 d = 0.76	Positive
Wiboolyasarin et al. (2024)	16.71	17.05	16.94	19.89	CG: η^2 = 0.055 EG: η^2 = 0.468	Pretest: p = 0.898 Posttest: p = 0.003	Positive
Zhang et al. (2023b)	/	/	EG1 = 11.66 EG2 = 10.57	EG1 = 31.03 EG2 = 28.29	/	/	Positive
Zhang et al. (2023c)	/	/	4.93	6.20	/	/	Positive

Five of the seven chatbot studies examined the use of ChatGPT, a chatbot currently receiving considerable scholarly attention, though they slightly differed in the intervention for the CG. Four studies (Boudouaia et al., 2024; Ghafouri et al., 2024; Mahapatra, 2024; Song & Song, 2023) involved an EG receiving ChatGPT-generated writing feedback and a CG receiving feedback from teachers. According to statistics in Table 10, these studies consistently showed improved writing scores in both groups from the pretest to the post-test, with the EG gaining significantly higher post-test scores than the CG. In the remaining study (Wiboolyasarin et al., 2024), the EG received personalised and timely feedback on three writing tasks from ChatGPT, while the CG simply completed the tasks without any feedback. Results revealed a significant gain in writing proficiency only in the EG, which generated a large effect size of AI feedback (η^2 = 0.468, cf. Table 10). However, due to the lack of feedback in

the CG, caution should be applied when attributing the writing improvement to the use of AI or the provision of feedback more generally. The study was thereby rated "moderate" RoB.

Two other chatbot studies by Zhang et al. (2023b; 2023c) investigated the impact of chatbots on EFL argumentative writing through one-group pretest-posttest studies. The chatbots in these studies featured interactive feedback, personalised learning, and multimedia instructional content, which were expected to improve writing skills. Although the participants scored higher in post-tests, the lack of a CG made it unclear whether the observed improvements were due to the AI intervention or natural progression. The weaknesses in study designs resulted in "moderate" RoB in these studies.

Table 11. Extracted data from writing studies on AI-assisted writing and evaluation tools

Reference	**CG**		**EG**		**Significance values and effect sizes**		**Result**
	Mean pre-test score	**Mean post-test score**	**Mean pre-test score**	**Mean post-test score**	**Pre/post-test**	**Between groups**	
Chang et al. (2021)	59.35	74.13	58.17	76.00	/	p = .044 d = .603	Positive
Gayed et al. (2022)	Mean Difference (Post-Pre): Lexical diversity: vocd-D = 2.298 MTLD = 1.3 Fluency: Production rate = 6.2 Syntactic complexity = 0.34				/	vocd-D: p = 0.47 MTLD: p = 0.81 Production rate: p = 0.38 Syntactic complexity: p = 0.0315 d = 1.05	Mixed
Hwang et al. (2023)	CG1 = 15.210 CG2 = 16.697	CG1 = 25.072 CG2 = 24.747	14.424	31.504	/	p < .001 η^2 = .142	Positive
Kostikova & Miasoiedova (2019)	/	/	3.05	4.15	/	/	Positive
Liu et al. (2023)	/	15.79	/	16.60	/	η^2= 0.040	Positive

Luo & Liang (2022)	72.13	78.34	72.70	83.38	/	Pretest: p = 0. 716 Posttest: p = 0.000	Positive
Shafiee Rad et al. (2023)	38.8	41.9	38.3	78.3	CG: d = .02 EG: d = .79	p = .000 η_p^2 = .157	Positive
Sun (2022)	/	/	54	78	p = 0.000	/	Positive
Wale (2024)	447.872	458.511	450.444	546.889	/	p < .05 d = 0.924	Positive
Wale & Kassahun (2024)	44.7	45.8	45.0	54.6	/	d = 0.9	Positive
Wang (2022)	/	/	7.28	8.76	p = 0.005 d = -0.56	/	Positive
Zhang (2024)	/	/	Mean Difference (Post-Pre): Content = 0.576 Organization = 0.334 Accuracy = 0.235		Content: p < .001 Organization: p < .001 Accuracy: p < .05	/	Positive

Of the 12 studies on AI-assisted writing and evaluation tools, six studies (Chang et al. 2021; Hwang et al., 2023; Luo & Liang, 2022; Shafiee Rad et al., 2023; Wale, 2024; Wale & Kassahun, 2024) adopted similar research designs to some of the above-mentioned chatbot studies (cf. Boudouaia et al., 2024; Ghafouri et al., 2024; Mahapatra, 2024; Song & Song, 2023), involving EGs receiving automated writing feedback from the AI systems and CGs receiving traditional feedback from teachers or peers. Consistent with findings in those chatbot studies, the EG's writing test scores were more significantly improved in the post-test.

Four other studies (Kostikova & Miasoiedova, 2019; Sun, 2022; Wang, 2022; Zhang, 2024) adopted one-group pretest-posttest designs, all demonstrating a notable gain in writing test scores after the intervention. Similar to the weakness pointed out in Zhang et al.'s (2023b; 2023c) studies, the absence of CG leads to problems with attributing the improvements to AI interventions.

Among the remaining two studies, Gayed et al. (2022) adopted a counter-balanced experimental design and compared the writings of two groups of participants. Both groups used AI KAKU (i.e., an AI-based writing assistant offering context-aware word suggestions and reverse translation) and Google Docs (i.e., a traditional word processor providing basic proofreading support such as spelling checks), but in a different order. It was found that the participants produced writing with higher levels of lexical diversity and fluency in the AI-assisted condition; yet, differences in writing quality between conditions failed to achieve the significance level except for the syntactic complexity

measure. The non-significance might be due to the participants' unfamiliarity with the AI tool, since they simply received a 5-minute training session on the tool's functions. Liu et al. (2023) studied the effect of incorporating a reflective thinking promotion mechanism into an AI-assisted writing system by dividing participants into an EG, who was encouraged to write reflections on AI-generated feedback, and a CG, who simply received feedback from the AI tool. Results showed that reflective thinking helped participants significantly improve their post-test writing performance, yielding an effect size of $\eta^2 = 0.040$ (see Table 11).

Despite the general consistency in pretest-posttest design and support for effectiveness, the writing studies varied in terms of specific evaluation rubrics for writing performance, which resembles the speaking studies. For example, several studies assessed writing performance according to different sub-measures (Boudouaia et al., 2024; Gayed et al., 2022; Hwang et al., 2023; Kostikova & Miasoiedova, 2019; Song & Song, 2023; Sun, 2022; Zhang, 2024), whereas the remaining studies reported overall writing scores (Chang et al., 2021; Ghafouri et al., 2024; Liu et al., 2023; Luo & Liang, 2022; Mahapatra, 2024; Shafiee Rad et al., 2023; Wale, 2024; Wale & Kassahun, 2024; Wang, 2022; Wiboolyasarin et al., 2024; Zhang et al., 2023b; Zhang et al., 2023c). The inconsistency in marking descriptors may limit the comparability of findings across studies. Furthermore, among the studies adopting sub-dimensions of writing performance, Gayed et al. (2022) observed significant improvements only in the syntactic complexity measure. Further investigations can explore whether AI interventions have a trade-off effect on various writing measures.

4.2.8 Evidence of learning outcomes studies on integrated language skills

8 studies examined the effectiveness of AI tools in enhancing integrated language skills (see Table 12).

Five of them compared AI-mediated instruction (EG) with traditional teacher-led instruction (CG), which covered various types of AI technologies, including AI-assisted language learning applications or platforms (Azamatova et al., 2023; Kim, 2022a; Wei, 2023), chatbots (Nozhovnik et al., 2023), and ITSs (Liu & Ardakani, 2022). For example, Kim (2022a) studied an AI-assisted language learning platform for TOEIC test preparation named Soljam. The platform could provide diagnostic tests, tailored instruction, interactive feedback, and individualised practice by tracking students 'learning progress. In Nozhovnik et al. (2023), the EG received instruction from a chatbot integrating gamification elements, automated feedback, and timely feedback. Liu and Ardakani (2022) examined their self-developed ITS which featured personalising learning contents according to students' emotional states. All studies, except Liu and Ardakani (2022), showed that both groups improved their test scores from the pretest to the post-test, with the EG exhibiting more gains. In Liu

and Ardakani (2022), the researchers observed that their system failed to exert a statistically significant influence on participants' language skills.

Table 12. Extracted data from studies on integrated language skills

Reference	CG		EG		Significance values and effect sizes		Result
	Mean pre-test score	**Mean post-test score**	**Mean pre-test score**	**Mean post-test score**	**Pre/post-test**	**Between groups**	
Azamatova et al. (2023)	15.97	25.66	15.78	29.13	/	$p < .001$	Positive
Kim (2022a)	Mean Difference (Post-Pre): Listening = 33.64 Reading = 10.76		Mean Difference (Post-Pre): AI CALL: Listening = 49.58 Reading = 27.32 AI MALL: Listening = 34.63 Reading = 14.58		CG: $p < .001$ AI CALL: $p < .001$ AI MALL: $p < .001$	Listening: $p = .014$ Reading: $p = .000$	Positive
Kim (2022b)	/	/	345.08	389.08	$p < .001$	/	Positive
Kim et al. (2022)	/	/	Mean Difference (Post-Pre): Cognitive: Grammar = 1.44 Vocabulary = 2.72 Emotional: Grammar = 0.3 Vocabulary = 3.54 Autonomy: Grammar (G) = 0.9 Vocabulary (V) = 6.88		Cognitive: G: $p < .05$ V: $p < .05$ Emotional: G: $t = 0.519$ V: $p < .05$ Autonomy: G: $t = 1.977$ V: $p < .01$	/	Positive
Liu & Ardakani (2022)	0.760	0.773	0.741	0.775	CG: $p = 0.752$ EG: $p = 0.281$	/	Neutral
Nozhovnik et al. (2023)	67.00	75.80	65.55	80.70	CG: $d = 3.09$ EG: $d = 11.3$	$p = .003$	Positive
Wei (2023)	44.39	61.11	43.21	73.86	/	$\eta^2 = 0.81$	Positive
Zhou (2023)	/	/	71.8	84.2	$p < .01$	/	Positive

Two studies involved intervention with one group. These cohort studies also demonstrated a positive effect of AI tools on enhancing language skills (Kim, 2022b; Zhou, 2023). Zhou (2023) studied the influence of an ML-powered learning platform on students' learning performance by comparing pre-intervention and post-intervention test scores. The platform, with its ability to tailor to students' learning needs through gamification and interactive activities, improved the participants' English test scores by 14.7% over six months. Kim (2022b) investigated the effect of Soljam on Korean university students' English learning over 10 weeks. Findings showed an overall significant improvement in all students' TOEIC scores, with a more notable improvement observed among female students. However, similar to the above-mentioned studies without CG, the study findings should be interpreted with caution, since the improvements might be ascribed to participants' natural progression in language proficiency.

In the remaining study, Kim et al. (2022) examined the impact of combining teacher support with an ITS on English learning. The researchers compared three types of teacher support: (1) cognitive support, where teachers provided explicit learning techniques and explanations; (2) emotional support, where teachers provided encouragement and positive feedback; and (3) autonomy support, where teachers encouraged students to manage their learning materials and progress. Results showed significant gains in vocabulary scores in all groups. Notable improvements in grammar scores were observed in all groups. However, only the cognitive support group demonstrated statistically significant gains, suggesting a need for collaboration between teachers and AI tools to optimise language learning outcomes. A limitation of the study was the lack of a CG who did not use the ITS, causing difficulties in isolating the effect of AI intervention.

In response to RQ2a, an overall 51 studies reported findings on AI-assisted language learning outcomes in higher education settings, which mostly substantiated the positive role of AI technologies as either alternatives or supplements to traditional instruction. However, the RoB in studies may limit the trustworthiness of findings. The cumulative confidence in findings across studies will be discussed in Section 5.2.1.

4.3 Evidence of learning perceptions

Following the synthesis of L2/FL acquisition outcomes, this section turns to an in-depth analysis of findings on learning perceptions to answer RQ2b. Learners' perceptions were mainly collected from questionnaires, interviews, and focus group data. The findings can be categorised into the following main themes: (1) general attitudes, (2) motivation or engagement, (3) willingness to communicate, (4) learning anxiety, and (5) learning interest. Each of the following subsections will expound on one of the themes.

4.3.1 General attitudes

41 studies examined participants' general attitudes towards the use of AI tools in language learning, identifying both affordances and challenges.

In terms of the benefits, most studies reported participants' perceived effectiveness and perceived ease of use of different types of AI technologies. For example, Shaikh et al.'s (2023) quantitative survey study revealed that the participants agreed on the effectiveness of ChatGPT in assisting English learning, and they reported high levels of ease of use (M = 43.30 out of 55) and usefulness (M = 30.80 out of 40) while using the tool for different learning tasks such as writing, grammar, and vocabulary practice. In two qualitative case studies, Y.-L. Chen et al. (2022) investigated the application of speech recognition tools among two Taiwanese students, while Oraif (2024) focused on ITS use among six female students in Saudi Arabia. Observation and interview data in both studies revealed that the users held positive attitudes towards the interactive learning experience offered by these AIEd tools. However, the small sample size may limit the generalisability of results and cause selection bias. Quantitative data from a larger population is thereby needed to complement the findings.

Only Zhang et al. (2023b; 2023c) found from questionnaire and interview data that the participants perceived chatbot-based learning as less effective for improving argumentative writing, despite a substantive increase in post-intervention writing scores (cf. Section 4.2.7). The result may be explained by the relatively short duration of intervention (5 weeks), which might be insufficient for participants to notice writing improvements and internalise the learning outcomes.

Problems with AIEd tools have also been identified. Firstly, some learners reported that these tools could not provide authentic and reliable interactions. For example, Tegos et al. (2014) examined a conversational agent designed to facilitate classroom communication. While questionnaire and focus group data indicated that participants generally perceived the system as effective in facilitating vocabulary development, some learners found its computer-generated voice unnatural. Similarly, three studies on ChatGPT (Karataş et al., 2024; Muniandy & Selvanathan, 2024; Xiao & Zhi, 2023) revealed participants' concerns about the reliability of ChatGPT-generated responses. Since these studies focused on ChatGPT-3.5, future research is needed to examine whether the use of ChatGPT-4 would generate different findings.

Secondly, studies reported personalisation issues. As H. Chen et al. (2020) noted, some participants complained that the chatbot in this study failed to tailor responses to their needs, although they demonstrated tangible vocabulary learning outcomes after using the AI tool (cf. Section 4.2.1). The repetitive practice questions provided by the chatbot also led to boredom among the participants.

Finally, some participants expressed privacy concerns about the AIEd tools regarding data collection and storage (Jamshed et al., 2024). Overall, learner perceptions of AI-assisted language learning tools were mixed.

4.3.2 Motivation

11 studies examined the effect of AI technologies on learning motivation, all reporting positive findings.

In studies comparing AI interventions (EG) against traditional controls (CG), the EG consistently exhibited higher motivation levels than the CG in post-intervention questionnaires and interviews (Azamatova et al., 2023; Kemelbekova et al., 2024; Nozhovnik et al., 2023; Song & Song, 2023; Wang & Feng, 2023; Wei, 2023; Wu, 2022). For example, Wang and Feng (2023)[95] studied the application of ChatGPT for reading assistance. The EG was assisted by ChatGPT for story background interpretation, vocabulary explanation, and sentence translation, while the CG employed electronic dictionaries to read traditional paper books. Questionnaire data showed that the EG reported higher reading speed, comprehension rate, and motivation levels. One limitation was the relatively short period of intervention (4 weeks), making it difficult to understand the long-term effect on motivation.

Studies that involved only one cohort group reported similar findings on enhanced motivation for language learning (Chew & Chua, 2020; Liao, 2023; Markus et al., 2023; Peña-Acuña & Crismán-Pérez, 2022). For instance, Chew and Chua (2020) examined the application of a humanoid robot for basic Chinese vocabulary learning. The robot, equipped with speech recognition techniques and adaptive learning algorithms, offers interactive feedback and personalised lessons based on students' performance. Interviews with six university students revealed that learning with the robot enhanced their motivation. The self-report data could be more robust if further substantiated by quantitative measures. Similarly, Peña-Acuña and Crismán-Pérez (2022) focused on Papua, an AI-assisted language learning application featuring voice recognition and personalised learning. Interview data also showed participants' improved motivation after using the application.

All these studies, except for Liao (2023), collected self-report motivation data from questionnaires and interviews only one time after the intervention. This approach potentially led to an overestimation of the effect of AIEd tools on motivation, since participants might report improved motivation shortly after using the tool due to novelty effects, but their enhanced motivation might not be sustained over time. To reduce these biases, future studies can combine self-report data with objective measures and collect data multiple times to provide more comprehensive findings on learning motivation.

4.3.3 Willingness to communicate

Five speaking studies investigated the effect of AI tools on willingness to communicate (WTC) in L2 or FL, all observing an improved WTC after the intervention.

Four studies compared EGs who practised speaking with chatbots (Fathi et al., 2024; Kim & Su, 2024) or AI-assisted language learning applications (Shafiee Rad, 2024; Zhang et al., 2024) against CGs who interacted with peers or teachers. These studies adopted validated questionnaires consisting of Likert-scale items to gauge participants' perceived WTC in various situations such as inside classrooms or with strangers. The results consistently revealed that the EGs exhibited significantly higher WTC levels than the CGs. Additionally, all studies, except Kim and Su (2024), reported a large effect size of the examined AI tools on WTC. However, since both CGs and EGs in these studies received regular teacher-led speaking instruction, it seemed challenging to isolate the potential effect of traditional teaching activities on WTC. To address confounding variables, further investigations could use the amount of time for teacher-led instruction as a covariate. Moreover, the geographical limitation of these studies to China and Iran restricted the generalisability of findings.

The remaining study, Ayedoun et al. (2019), examined a chatbot equipped with communicative strategies (e.g., requests for clarification) and affective backchannels (e.g., verbal and non-verbal cues for encouragement such as nodding). Researchers found from self-report survey data that combining communicative strategies with affective backchannels was most effective for improving participants' WTC.

Overall, studies demonstrated the benefits of AI tools in enhancing WTC, but potential biases from confounding variables and selection bias necessitate cautious interpretation of the findings. Studies with more rigorous controls are needed to assess the effect of AI-assisted applications on L2/FL WTC in different countries.

4.3.4 Learning anxiety

Five studies used questionnaires and interviews to examine the effect of AI tools on learning anxiety, yielding inconsistent results.

Three studies compared AI-assisted EGs with CGs receiving traditional instruction. Kim and Cha (2023) investigated the influence of AI translators on reading comprehension. Apart from findings on learning outcomes (cf. Section 4.2.6), the study identified from open-ended questionnaires that nearly 20% of the EG regarded AI translators as anxiety-relievers in reading. Similarly, Kim and Su (2024) and Zhang et al. (2024) established the effectiveness of chatbots in relieving speaking anxiety from Likert-scale questionnaires. A shared limitation of these studies was

the lack of significance values and effect size reporting, which rendered the conclusions less convincing.

By contrast, the remaining two studies (Çakmak, 2022; El Shazly, 2021) reported increased speaking anxiety from scale questionnaires, although they observed participants' improved speaking performance after using chatbots (cf. Section 4.2.5). The result might arise from problems with the chatbot to recognise certain sounds in students' oral output and the ensuing communication breakdowns that frustrated the students. According to the Interaction Hypothesis, communication breakdowns on the one hand bring opportunities for corrective feedback and input modifications, on the other hand, it might increase anxiety when students struggle to be understood by the AI tool. Therefore, it is essential for language learners to manage the benefits and drawbacks of communication breakdowns.

4.3.5 Learning interest

Four studies focused on the effect of chatbots on learning interests, generating mixed results.

In the study mentioned above by Chew and Chua (2020) (cf. Section 4.3.2), interview data also showed that participants enhanced their interest in learning after the intervention. Similar results on enhanced learning interest were observed in Wang and Feng's (2023) reading study (cf. Section 4.3.2). However, since these studies either did not specify the duration of the intervention or lasted relatively short, it should be cautioned that the positive effect on learning interest might arise from the novelty effect of newly introduced technologies.

Two speaking studies by Fryer et al. (2017, 2019) examined the effect of chatbots on long-term learning interest to further explore whether the novelty effect could be sustained over time. Fryer et al. (2017) compared chatbot against human task partners to investigate their effect on 122 Japanese EFL university students' speaking task interest. Quantitative results from Likert-scale questionnaire items showed a significant drop in the chatbot group's task interest over time, while no decline in the human partner group, suggesting that the novelty effect did not occur in the human partner group. Another study by Fryer et al. (2019) also compared the chatbot intervention against a human partner control. Quantitative survey data revealed that participants' interest in chatbot conversation decreased because of the diminishing novelty effect, but it showed a notable rebound after five months. The results suggested the chatbot's potential to stimulate and maintain interest in the long term, despite an initial decline. The aforementioned studies examined ITS and chatbots. Future studies can investigate the effect of other AI tools on language learning interests.

To answer RQ2b, 59 of the 71 studies in tertiary education settings reported findings on learner perceptions, yielding mixed results. Problems with study designs, such as limited intervention

durations and the introduction of confounding variables, added further caution regarding the reliability of findings. These concerns will be examined in the following Discussion section.

Chapter 5 Discussion

This section first discusses the general characteristics of the 105 studies meeting the inclusion criteria and then conducts a detailed discussion of the main findings on AI-assisted language learning outcomes and learning perceptions in tertiary settings.

5.1 RQ1: What is the extent of research that has been conducted?

RQ1 concerned the extent of extant research on AI-assisted language learning. Overall, the review included 105 studies published since 1956, when the concept of AI was proposed, with the first eligible study published in 1995. The synthesis in Section 4.1 highlights the following characteristics of the included studies: (1) a notable rise in the number of publications, especially in the past four years; (2) a predominance of research in higher education settings, particularly among ESL/EFL learners with diverse language backgrounds; (3) a wide geographical distribution of studies, with primary interests in Asian countries; (4) the employment of various AIEd technologies, with chatbots receiving the most research attention; (5) a strong focus on using AI tools to develop writing, speaking, and integrated language skills; (6) a majority of mixed methods studies; (7) a frequent absence of theoretical framework guiding the research. The remainder of this section discusses these research trends.

Regarding the annual distribution of publications, the recent growth in the number of studies is unsurprising, given the burgeoning use of AI technologies to empower educational practices. The publication trend aligned with findings from earlier systematic reviews on AIEd (X. Chen et al., 2020; Zawacki-Richter et al., 2019) and AILEd (Huang et al., 2023; Liang et al., 2023). Driven by ongoing technological breakthroughs in the Transformer algorithms (Law, 2024), research output in AI-assisted language learning is expected to grow in the coming years.

In terms of educational levels, the review was consistent with previous studies (Liang et al., 2023; Yang & Kyun, 2022) in identifying higher education as the primary research context. This focus might be attributed to the convenience of accessing tertiary-level students who were cognitively mature enough to harness advanced AI technologies and complete relevant tasks. However, it is worth noting that AI technologies have been increasingly applied in K-12 education (UNESCO, 2022a). Taking China, the country of primary focus, as an example, several policies (e.g., “Opinions on the Implementation of Project 2.0 to Improve the Application Capability of Information Technology for Elementary and Secondary School Teachers”) have encouraged teachers to employ AI technologies to assist effective teaching and learning (Xing, 2023). These supportive policies and developments in

AI technologies indicate a need for research on the under-explored primary and secondary school students.

Other information about participant characteristics, including gender and L2 proficiency, was comparatively less explicitly reported across studies. Nevertheless, several included studies have examined the role of gender (Kim, 2022)[16] and language levels (Kim et al., 2021a; Kim, 2022; Liao, 2023)[16, 61, 95] in students' AI-assisted language learning outcomes. In terms of gender, Kim (2022) studied the effect of an AI-assisted TOEIC learning platform among Korean EFL students. The researcher found significant differences in TOEIC listening and reading scores by gender, suggesting a need to specify gender information in future studies. Regarding language proficiency, while Kim (2022) identified no significant differences among learners of different language levels, Kim et al. (2021a)[95] observed significant differences in the intonation and stress scores between low and intermediate language learners after using chatbots for TOEIC speaking practice. Due to the inconsistent findings on the influence of language levels on AI intervention effectiveness, future studies should specify participants' L2 proficiency. Additionally, although some studies categorised their participants by L2 levels, they did not use standardised test scores as a basis for classification. It is expected that future studies report language proficiency based on standardised test results, such as IELTS and TOEIC, to enhance the comparability of findings across studies.

Regarding the research contexts, the finding that most studies were conducted in Asian contexts diverged from Huang et al. (2023) and Liang et al. (2023), who identified the US as the major country of contribution. This discrepancy might be due to the recent rapid integration of technology into classrooms in Asian regions (UNESCO, 2022b). Studies in other regions are warranted to explore whether the location of study could influence research findings on the effectiveness of AI. Additionally, consistent with Yang and Kyun's (2022) review finding, most studies focused on learning English as an FL, with only sporadic studies conducted in other language learning contexts. This trend underscores a need for studies focusing on the acquisition of languages other than English to provide a comprehensive understanding of the effectiveness of AI-assisted language learning across various research contexts.

In terms of studied language skills, the current review reported that writing received particular research attention, which supported what was found in Huang et al. (2023) and Yang and Kyun (2022). Additionally, speaking and vocabulary studies were prominent in the number of publications, which slightly differed from Liang et al. (2023), who found that writing, reading, and vocabulary were the top three most-researched language skills. The increasing research interest in speaking might be attributed to the recent innovations and developments in chatbots, especially the iterative GPT series (Gillani et al., 2023). The rapid advancements in chatbots have also been manifested in the

largest proportion of chatbot studies in this review and the recent surge in the number of publications, driven by the emergence of LLM-powered chatbots such as ChatGPT.

Regarding research designs, most studies adopted a mixed-methods approach, differing from previous reviews (Liang et al., 2023; Yang & Kyun, 2022) which found that quantitative methods were most frequently employed. By combining quantitative and qualitative methods, researchers can gain a comprehensive understanding of learning outcomes from pretest-posttest measures and questionnaires, while also gathering learning perception findings from interviews. Given that most studies were (quasi-)experimental, future research can be conducted in authentic classroom settings to evaluate the practical effectiveness of AI interventions.

As for theoretical backgrounds, the vast majority of studies did not state this information, which aligned with the finding of X. Chen et al. (2020) and Wang et al. (2024)**.** Among the remaining studies, the most frequently used theory was Vygotsky's (1978) SCT, which underlines the role of social interactions in learning. A core concept of the theory is ZPD, which refers to the tasks that can be performed under guidance but cannot be performed independently. AI applications function to provide such scaffolded support and meet personalisation needs (Qiao & Zhao, 2023). Another prominent perspective was the cognitivist approach, including the Input Hypothesis, the Interaction Hypothesis, and the Noticing Hypothesis. Given insufficient input in traditional EFL classrooms, AIEd technologies can address this problem by offering personalised comprehensible input and encouraging learners to engage in interactive learning activities (see Section 2.3.1 for a detailed explanation). Altogether, the cognitivist and sociocultural perspectives on SLA both lend support for the use of AI in language learning.

5.2 RQ2: What is the extent of empirical evidence for the effectiveness of AI-assisted language learning?

RQ2 explored the extent of empirical evidence of AI-assisted language learning outcomes and learning perceptions in higher education. The phase of higher education was chosen to make the project manageable while including as much research as possible for a comprehensive review. The author recognised the selection bias for only synthesising tertiary-level studies due to time constraints.

5.2.1 The extent of findings on learning outcomes

A total of 51 studies reported findings on substantive learning outcomes, consisting of 36 studies adopting pretest-posttest control group designs, 14 adopting one-group pretest-posttest designs, and

one (Gayed et al., 2022) (cf. Section 4.2.7) adopting the counter-balanced research design. The study results were labelled “positive”, “negative”, “neutral”, or “mixed”.

To illustrate AI-assisted language learning outcomes across different language skills, Figure 13 was created using Python (Plotly Technologies Inc., 2015). The figure shows that studies with rigorous study designs (i.e., rated “low” RoB) reported positive learning outcomes in all grammar studies and some of the studies on vocabulary, speaking, writing, and integrated skills. However, more speaking and writing studies, represented by larger bubbles, produced positive results but received “moderate” RoB ratings, suggesting that their findings should be interpreted with caution. Additionally, the robustness of positive results in pronunciation and listening was affected by moderate or high RoB. Taken together, while 47 studies showed positive learning outcomes after AI interventions, the cumulative weight of evidence was limited due to RoB in individual studies. Following the GRADE approach, the reviewers rated the cumulative quality of learning outcome evidence as “moderate”.

Figure 13. Reported learning outcomes and RoB

Studies reporting negative, mixed, and neutral learning outcomes are discussed below. According to Section 4.2.4, Li and Peng (2021)[86] [2] was the only study yielding negative results. The researchers observed lower post-intervention listening scores in both CG and EG, with no significant differences between groups. This result might be attributed to the mismatched difficulty levels of the

[2] The superscript number represents the study ID in Appendix 5.

pretest and post-test. Due to a lack of information about test design procedures and specific test items, it becomes uncertain whether the study ensured that the pretest and post-test were of similar difficulty levels. Additionally, the study involved only 59 freshmen from a single university, which might limit the generalizability of findings. The moderate RoB (see Figure 13) stemmed from a lack of reliable outcome measures and participant selection bias, suggesting that future studies can be conducted among a larger sample and be specific about the measurements to ensure the comparability of difficulty levels of the pretest and the post-test.

Gayed et al. (2022)[79] was the only study employing a counter-balanced experimental design to explore the influence of AI writing tools on the lexical diversity and fluency of writing, reporting mixed findings for different dimensions of writing. As mentioned in Section 4.2.7, the inconsistencies among measures for different descriptors suggest a need to specify the marking schemes, report scores for each descriptor, and explore the trade-off effects among different aspects.

Two studies failed to find evidence of differences. The first study by Kim and Cha (2023)[63] compared three groups who read texts with one of (i) AI translators, (ii) revised translations from AI translators, and (iii) no AI translators. The researchers identified an improvement in reading comprehension test scores within all three groups, but no significant differences were found among groups. The findings suggested that while the improvement could not be conclusively attributed to the AI intervention due to the lack of significance, the use of AI did not prevent language acquisition. This goes against Garcia and Pena (2011) and Karnal and Pereira (2015), who pointed out the risks of fewer learning outcomes when MT was used. Additionally, the non-significant difference between the AI translation-assisted group and the revised AI translation-assisted group was inconsistent with Niño (2008), who found notable benefits of revised translation in improving writing outcomes. This may reflect the improvements in MT performance over the years (Lee, 2023) (see Section 2.3.2 for details), which provides contextually appropriate translations and reduces the need for human post-editing.

The second study (Liu & Ardakani, 2022)[69] investigated the effectiveness of a researcher-designed ITS and identified statistically non-significant improved test scores in both the AI-assisted EG and the CG receiving traditional teacher-led instructions. While the researchers suggested that the non-significant differences in test scores imply a need to further develop the intelligent system, the result may also be attributed to inherent methodological flaws, such as the failure to ensure comparability of groups before interventions, which led to a classification of “moderate” RoB in this study.

Overall, the analysis of cumulative confidence in findings suggested that although most studies claimed that AI-assisted language learning was effective, robust evidence for positive learning outcomes was limited.

5.2.2 The extent of findings on learning perceptions

Regarding learning perceptions, the considerable number of publications (n = 59) reflected the research trend identified by Yang and Kyun (2022) that learning attitudes have garnered increasing scholarly attention. Findings on learning perceptions indicated both opportunities and challenges of AI technologies, conforming to what has been summarised in previous systematic reviews (Liang et al., 2023). The following contents will discuss findings on positive and negative learning perceptions.

Opportunities included general positive attitudes towards AI-supported pedagogies, stronger motivation for language learning, and enhanced willingness to communicate, which have been reported across studies on all five types of AI tools. However, methodological flaws, such as short intervention durations, over-reliance on self-report data, and confounding variables, highlight a need for more rigorously designed future research.

Challenges encompassed technical and ethical issues. The technical issues, primarily presented in chatbot studies, included the inaccuracies in speech recognition and unreliability of speech output, which caused anxiety and loss of interest among language learners. These problems, consistent with previous findings (Coniam, 2014; McCrocklin, 2016), demonstrated a need to train chatbots on non-native speakers' oral input, despite the recent technological breakthroughs in ASR and chatbots (Han, 2024; Hwang & Chang, 2023). Ethical issues, such as privacy concerns, echoed findings from prior studies (Bond et al., 2024; Gillani et al., 2023). Altogether, the mixed learning perception findings suggested directions for future research and technological improvements.

6 Conclusion

6.1 Summary of findings

This study systematically reviewed publications on the effectiveness of AI-assisted tools in L2/FL learning. Regarding the extent of existing research (RQ1), the systematic review searched within 9 electronic databases and identified 105 eligible studies, involving 10 kinds of L2/FL, 23 locations of study, 5 types of AI technologies, 8 language skills, and participants from diverse gender, age, and proficiency groups. Specifically, the review revealed research trends in implementing chatbots to develop writing skills among tertiary-level students in Asian countries, such as China and Korea. A common problem with the retrieved studies, however, was a lack of theoretical backgrounds.

Due to the larger-than-expected number of studies meeting the inclusion criteria and the dominance of research in higher education contexts, RQ2 was narrowed down to delve into higher education settings and report evidence of learning outcomes and learning perceptions. The synthesis of 71 studies in tertiary contexts found that 51 reported findings on learning effectiveness and 59 included findings on learning perceptions. Positive learning outcomes of AI technologies were generally reported across studies on different language skills, except for four studies. However, the moderate RoB across studies limited the robustness of conclusions. Additionally, studies on learning perceptions reported generally consistent results on learning motivation and willingness to communicate yet mixed findings on learning anxiety and learning interest. Given the opportunities and challenges of AI-assisted language learning, the subsequent sections present the implications of the findings and critically discuss the limitations of this review.

6.2 Theoretical and practical implications

The review findings suggested implications for language learners, teachers, researchers, and AI developers. Practically, given the effectiveness of AI-assisted language learning tools, language learners and teachers can integrate the novel technology into the learning and teaching process to enhance learning outcomes. Developers are also guided to address the problems with existing AI tools.

Theoretically, the reported effectiveness of AI-assisted language learning supports the Input Hypothesis, the Interaction Hypothesis, and the Sociocultural Theory in SLA. Considering that most studies lacked theoretical frameworks, it is suggested that future research and study designs can be underpinned by SLA or technology-related theories.

6.3 Limitations and future research

A major limitation of this review concerns its selection bias, since it only conducted an in-depth review of studies in higher education contexts due to time constraints. Future systematic reviews should synthesize studies in all educational phases to provide a more comprehensive picture of research trends. If focusing on a specific educational stage, future reviews are recommended to incorporate search terms related to "higher education", such as university OR college OR "tertiary education", to enhance the relevance of retrieved studies.

Another limitation is this review excluded studies on teachers' application of AI educational tools to maintain a manageable amount of data. Future empirical studies and systematic reviews can explore the similarities and differences in the perceptions of using AI tools in language learning between learners and teachers.

The third limitation is the review was limited to publications in English and Chinese, which might affect the comprehensiveness of this review. Future reviews can involve reviewers with diverse language backgrounds to retrieve data from more databases and examine whether any general characteristics of existing studies, such as research locations, would be different.

Finally, while most studies claimed the effectiveness of AI-assisted language learning, robust evidence remained limited. Future studies should employ more rigorous research designs to enhance confidence in attributing positive language learning outcomes and perceptions to AI interventions. The study bias can be addressed by controlling for confounding variables, considering the blinding of assessors, adopting reliable and transparent outcome measures, and reporting effect sizes. Furthermore, the credibility of the findings in this review can be strengthened if detailed justifications for the RoB ratings of each study are provided. This approach will guide directions for methodological improvements in future research.

References

* denotes that the paper was identified through the systematic search and included in the keyword map.

** denotes that the paper was identified through the systematic search and reviewed in depth.

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications*, *2*, 1-18. https://doi.org/10.1016/j.mlwa.2020.100006

**Agnes, D., & Srinivasan, R. (2024). Fostering Vocabulary Memorization: Exploring the Impact of AI-Generated Mnemonic Keywords on Vocabulary Learning Through Anki Flashcards. *World Journal of English Language*, *14*(2), 434-451. https://doi.org/doi:10.5430/wjel.v14n2p434

*Al Mahmud, F. (2023). Investigating EFL Students' Writing Skills Through Artificial Intelligence: Wordtune Application as a Tool. *Journal of Language Teaching and Research*, *14*(5), 1395-1404. https://doi.org/10.17507/jltr.1405.28

Amaral, L., Meurers, D., & Ziai, R. (2011). Analyzing learner language: Towards a flexible natural language processing architecture for intelligent language tutors. *Computer Assisted Language Learning*, *24*(1), 1–16. https://doi.org/10.1080/09588221.2010.520674

**Ayedoun, E., Hayashi, Y., & Seta, K. (2019). Adding Communicative and Affective Strategies to an Embodied Conversational Agent to Enhance Second Language Learners' Willingness to Communicate. *International Journal of Artificial Intelligence in Education*, *29*(1), 29–57. https://doi.org/10.1007/s40593-018-0171-6

**Azamatova, A., Bekeyeva, N., Zhaxylikova, K., Sarbassova, A., & Ilyassova, N. (2023). 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. *International Journal of Education in Mathematics, Science, and Technology*, *11*(6), 1458-1475. https://doi.org/10.46328/ijemst.3712

Baker, T., & Smith, L. (2019). *Educ-AI-tion Rebooted? Exploring the future of artificial intelligence in schools and colleges*. https://www.nesta.org.uk/report/education-rebooted/

*Bao, M. (2019). Can Home Use of Speech-Enabled Artificial Intelligence Mitigate Foreign Language Anxiety -- Investigation of a Concept. *Arab World English Journal*, *5*, 28-40. https://doi.org/10.24093/awej/call5.3

Boland, A., Cherry, G., & Dickson, R. (Eds.). (2017). *Doing a Systematic Review: A Students' Guide* (2nd ed.). SAGE.

Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, *21*(4), 1-41. https://doi.org/10.1186/s41239-023-00436-z

**Boudouaia, A., Mouas, S., & Kouider, B. (2024). A Study on ChatGPT-4 as an Innovative Approach to Enhancing English as a Foreign Language Writing Learning. *Journal of Educational Computing Research*, 1-29. https://doi.org/10.1177/07356331241247465

Briggs, N. (2018). Neural machine translation tools in the language learning classroom: Students' use, perceptions, and analyses. *JALT CALL Journal*, *14*(1), 3–24.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R.,

Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901. http://arxiv.org/abs/2005.14165

**Çakmak, F. (2022). Chatbot-Human Interaction and Its Effects on EFL Students' L2 Speaking Performance and Anxiety. *Novitas-ROYAL (Research on Youth and Language)*, *16*(2), 113-131.

Carbonell, J. R. (1970). AI in CAI: An artificial-intelligence approach to computer-assisted instruction. *IEEE Transactions on Man Machine Systems*, *11*(4), 190–202. https://doi.org/10.1109/TMMS.1970.299942

Chalmers, H., Brown, J., & Koryakina, A. (2023). Topics, publication patterns, and reporting quality in systematic reviews in language education. Lessons from the international database of education systematic reviews (IDESR). *Applied Linguistics Review*, 1-25. https://doi.org/10.1515/applirev-2022-0190

**Chang, T.-S., Li, Y., Huang, H.-W., & Whitfield, B. (2021). Exploring EFL Students' Writing Performance and Their Acceptance of AI-based Automated Writing Feedback. In *2021 2nd International Conference on Education Development and Studies* (pp. 31-35). ACM. https://doi.org/10.1145/3459043.3459065

*Chen, B., Bao, L., Zhang, R., Zhang, J., Liu, F., Wang, S., & Li, M. (2024). A Multi-Strategy Computer-Assisted EFL Writing Learning System With Deep Learning Incorporated and Its Effects on Learning: A Writing Feedback Perspective. *Journal of Educational Computing Research*, *61*(8), 60-102. https://doi.org/10.1177/07356331231189294

**Chen, H., Widarso, G., & Sutrisno, H. (2020). A ChatBot for Learning Chinese: Learning Achievement and Technology Acceptance. *Journal of Educational Computing Research*, *58*(6), 1161-1189. https://doi.org/doi:10.1177/0735633120929622

Chen, X., Xie, H., Zou, D., & Hwang, G.-J. (2020). Application and theory gaps during the rise of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, *1*, 1-20. https://doi.org/10.1016/j.caeai.2020.100002

Chen, X., Zou, D., Xie, H., Cheng, G., & Liu, C. (2022). Two Decades of Artificial Intelligence in Education: Contributors, Collaborations, Research Topics, Challenges, and Future Directions. *Educational Technology & Society*, *25* (1), 28-47. https://www.jstor.org/stable/48647028

**Chen, Y.-L., Hsu, C.-C., Lin, C.-Y., & Hsu, H.-H. (2022). Robot-Assisted Language Learning: Integrating Artificial Intelligence and Virtual Reality into English Tour Guide Practice. *Education Sciences*, *12*(7), 1-20. https://doi.org/10.3390/educsci12070437

Cherry, M. G., & Dickson, R. (2017). Defining My Review Question and Identifying Inclusion and Exclusion Criteria. In A. Boland, M. G. Cherry, & R. Dickson (Eds.), *Doing a Systematic Review: A Student's Guide* (2nd ed.) (pp. 68-86). SAGE.

**Chew, E., & Chua, X. N. (2020). Robotic Chinese language tutor: personalising progress assessment and feedback or taking over your job? *On the Horizon*, *28*(3), 113-124. https://doi.org/10.1108/OTH-04-2020-0015

*Chia, C.-L., & Li, C.-H. (2020). Artificial Intelligence in Practice: Conversation-Based Intelligent Tutoring System Conducts Research on Japanese Remedial Teaching. *Journal of Educational Practice and Research*, *33* (2), 1-42.

*Chien, Y.-C., Wu, T.-T., Lai, C.-H., & Huang, Y.-M. (2022). Investigation of the Influence of Artificial Intelligence Markup Language-Based LINE ChatBot in Contextual English Learning. *Frontiers in Psychology*, *13*, 1-8. https://doi.org/10.3389/fpsyg.2022.785752

Chiu, M.-C., Hwang, G.-J., Hsia, L.-H., & Shyu, F.-M. (2024). Artificial intelligence-supported art education: A deep learning-based system for promoting university students' artwork appreciation and painting outcomes. *Interactive Learning Environments*, *32*(3), 824–842. https://doi.org/10.1080/10494820.2022.2100426

Coniam, D. (2014). The linguistic accuracy of chatbots: Usability from an ESL perspective. *Text & Talk*, *34*(5), 545-567. https://doi.org/10.1515/text-2014-0018

Corbett, A. T., Koedinger, K. R., & Anderson, J. R. (1997). Intelligent tutoring systems. In M. G. Helander, T. K. Landauer, & P. V. Prabhu (Eds.), *Handbook of human-computer interaction* (2nd ed., pp. 849–874). Elsevier Science & Technology.

*Darejeh, A., Moghadam, T. S., Delaramifar, M., & Mashayekh, S. (2024). A Framework for AI-Powered Decision Making in Developing Adaptive e-Learning Systems to Impact Learners' Emotional Responses. In *2024 11th International and the 17th National Conference on E-Learning and E-Teaching* (pp. 1-6). IEEE. https://doi.org/10.1109/ICeLeT62507.2024.10493103

*Davis, G. M. (2022). *Humanlike Conversational Artificial Intelligence Agents for Foreign or Second Language Learning: User Perceptions, Expectations, And Interactions with Agents* (Publication Number 29342315) [Doctoral dissertation, Stanford University]. ProQuest Dissertations and Theses Global.

*Dhanapal, C., Asharudeen, N., & Alfaruque, S. Y. (2024). Impact of Artificial Intelligence Versus Traditional Instruction for Language Learning: A Survey. *World Journal of English Language*, *14*(2), 182-193. https://doi.org/10.5430/wjel.v14n2p182

Dickson, R., Cherry, M. G., & Boland, A. (2017). Carrying Out a Systematic Review as a Master's Thesis. In A. Boland, M. G. Cherry, & R. Dickson (Eds.), *Doing a Systematic Review: A Student's Guide* (2nd ed.) (pp. 20-45). SAGE.

Dundar, Y., & Fleeman, N. (2017). Developing My Search Strategy. In A. Boland, M. G. Cherry, & R. Dickson (Eds.), *Doing a Systematic Review: A Student's Guide* (2nd ed.) (pp. 86-106). SAGE.

**Duong, T., & Suppasetseree, S. (2024). The Effects of an Artificial Intelligence Voice Chatbot on Improving Vietnamese Undergraduate Students' English Speaking Skills. *International Journal of Learning, Teaching and Educational Research*, *23*(3), 293-321. https://doi.org/10.26803/ijlter.23.3.15

**El Shazly, R. (2021). Effects of artificial intelligence on English speaking anxiety and speaking performance: A case study. *Expert Systems*, *38*(3), 1-15. https://dx.doi.org/10.1111/exsy.12667

Ellis, R., & Barkhuizen, G. (2005). *Analysing Learner Language*. Oxford University Press.

**El-Zeiny, M. E., Faqeeh, M. H., Almahdawi, A., Abdo, M. A.-F., Mostafa, I., & Alahbabi, A. (2023). Enhancing Non-Native Speakers' Pronunciation: AI-driven Storytelling with Arabic Emphatic Consonants. In SNAMS 2023 General Chairs (Ed.), *2023 Tenth International Conference on Social Networks Analysis, Management and Security* (pp. 420-427). IEEE. https://doi.org/10.1109/SNAMS60348.2023.10375436

Esit, Ö. (2011). Your verbal zone: An intelligent computer-assisted language learning program in support of Turkish learners' vocabulary learning. *Computer Assisted Language Learning*, *24*(3), 211–232. https://doi.org/10.1080/09588221.2010.538702

Evers, K., & Chen, S. (2022). Effects of an automatic speech recognition system with peer feedback on pronunciation instruction for adults. *Computer Assisted Language Learning*, *35*(8), 1869–1889. https://doi.org/10.1080/09588221.2020.1839504

**Fathi, J., Rahimi, M., & Derakhshan, A. (2024). Improving EFL learners' speaking skills and willingness to communicate via artificial intelligence-mediated interactions. *System*, *121*, 1–17. https://doi.org/10.1016/j.system.2024.103254

*Feng, Y., & Wang, X. (2023). A comparative study on the development of Chinese and English abilities of Chinese primary school students through two bilingual reading modes: human-AI robot interaction and paper books. *Frontiers in Psychology*, *14*, 1-17. https://doi.org/10.3389/fpsyg.2023.1200675

Fleeman, N., & Dundar, Y. (2017). Data Extraction: Where Do I Begin? In A. Boland, M. G. Cherry, & R. Dickson (Eds.), *Doing a Systematic Review: A Student's Guide* (2nd ed.) (pp. 121-137). SAGE.

Frey, B. (Eds.). (2018). *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. SAGE.

**Fryer, L. K., Ainley, M., Thompson, A., Gibson, A., & Sherlock, Z. (2017). Stimulating and sustaining interest in a language course: An experimental comparison of Chatbot and Human task partners. *Computers in Human Behavior*, *75*, 461–468. https://doi.org/10.1016/j.chb.2017.05.045

**Fryer, L. K., Nakao, K., & Thompson, A. (2019). Chatbot learning partners: Connecting learning experiences, interest and competence. *Computers in Human Behavior*, *93*, 279-289. https://doi.org/10.1016/j.chb.2018.12.023

**Fu, S., Gu, H., & Yang, B. (2020). The affordances of AI-enabled automatic scoring applications on learners' continuous learning intention: An empirical study in China. *British Journal of Educational Technology*, *51*(5), 1674–1692. https://doi.org/10.1111/bjet.12995

Garcia, I., & Pena, M. I. (2011). Machine translation-assisted language learning: Writing for beginners. *Computer Assisted Language Learning*, *24*(5), 471–487. https://doi.org/10.1080/09588221.2011.582687

Gass, S. M. (1997). *Input, interaction, and the second language learner*. Lawrence Erlbaum Associates Publishers.

Gaudioso, E., Montero, M., & Hernandez-del-Olmo, F. (2012). Supporting teachers in adaptive educational systems through predictive models: A proof of concept. *Expert Systems with Applications*, *39*(1), 621–625. https://doi.org/10.1016/j.eswa.2011.07.052

**Gayed, J. M., Carlon, M. K. J., Oriola, A. M., & Cross, J. S. (2022). Exploring an AI-based writing Assistant's impact on English language learners. *Computers and Education: Artificial Intelligence*, *3*, 1-7. https://doi.org/10.1016/j.caeai.2022.100055

*Ghafouri, M. (2024). ChatGPT: The catalyst for teacher-student rapport and grit development in L2 class. *System*, *120*, 1-12. https://doi.org/10.1016/j.system.2023.103209

**Ghafouri, M., Hassaskhah, J., & Mahdavi-Zafarghandi, A. (2024). 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. *Language Teaching Research*, 1-23. https://doi.org/10.1177/13621688241239764

Gillani, N., Eynon, R., Chiabaut, C., & Finkel, K. (2023). Unpacking the "Black Box" of AI in Education. *Educational Technology & Society*, *26*(1), 99-111. https://doi.org/10.30191/ETS.202301_26(1).0008

*Gómez-Ortiz, R. E., & Buentello-Montoya, D. A. (2022). A Machine Learning-based Application to Practice Foreign Languages Using Images and Photographs. In C. Ceballos (Ed.), *2022 IEEE*

International Conference on Teaching, Assessment and Learning for Engineering (pp. 362-366). IEEE. https://doi.org/10.1109/TALE54877.2022.00066

*Guo, X., & Wen, Y. (2023). AI-powered Collaborative Activities for Chinese Vocabulary Learning. In J.-L. Shih, A. Kashihara, W. Chen, & H. Ogata (Eds.), *Proceedings of the 31st International Conference on Computers in Education* (pp. 819-824). Asia-Pacific Society for Computers in Education. https://hdl.handle.net/10497/27471

Han, Z. (2024). Chatgpt in and for second language acquisition: A call for systematic research. *Studies in Second Language Acquisition*, *46*(2), 301–306. https://doi.org/10.1017/S0272263124000111

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.

Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., & Vedel, I. (2018). *Mixed methods appraisal tool (MMAT), version 2018*. Registration of Copyright (#1148552), Canadian Intellectual Property Office, Industry Canada. http://mixedmethodsappraisaltoolpublic.pbworks.com/w/file/fetch/127916259/MMAT_2018_c rite ria-manual_2018-08-01_ENG.pdf

Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., & Vedel, I. (2020). *Reporting the results of the MMAT (version 2018)*. http://mixedmethodsappraisaltoolpublic.pbworks.com/w/file/fetch/140056890/Reporting%20t he%20results%20of%20the%20MMAT.pdf

Hooshyar, D., Ahmad, R. B., Yousefi, M., Fathi, M., Horng, S.-J., & Lim, H. (2016). Applying an online game-based formative assessment in a flowchart-based intelligent tutoring system for improving problem-solving skills. *Computers & Education*, *94*, 18–36. https://doi.org/10.1016/j.compedu.2015.10.013

**Hsu, M.-H., Chen, P.-S., & Yu, C.-S. (2023). Proposing a task-oriented chatbot system for EFL learners speaking practice. *Interactive Learning Environments*, *31*(7), 4297-4308. https://doi.org/10.1080/10494820.2021.1960864

*Hsu, T.-C., Chang, C., & Jen, T.-H. (2023). Artificial Intelligence image recognition using self-regulation learning strategies: effects on vocabulary acquisition, learning anxiety, and learning behaviours of English language learners. *Interactive Learning Environments*, 1-19. https://doi.org/10.1080/10494820.2023.2165508

**Huang, T., & Wang, L. (2023). Artificial intelligence learning approach through total physical response embodiment teaching on French vocabulary learning retention. *Computer Assisted Language Learning*, *36*(8), 1608-1632. https://doi.org/10.1080/09588221.2021.2008980

Huang, X., Zou, D., Cheng, G., Chen, X., & Xie, H. (2023). Trends, Research Issues and Applications of Artificial Intelligence in Language Education. *Educational Technology & Society*, *26*(1), 112-131. https://doi.org/10.30191/ETS.202301_26(1).0009

Hwang, G.-J., & Chang, C.-Y. (2023). A review of opportunities and challenges of chatbots in education. *Interactive Learning Environments*, *31*(7), 4099–4112. https://doi.org/10.1080/10494820.2021.1952615

**Hwang, W.-Y., Nurtantyana, R., Purba, S. W. D., Hariyanti, U., Indrihapsari, Y., & Surjono, H. D. (2023). AI and Recognition Technologies to Facilitate English as Foreign Language Writing for Supporting Personalization and Contextualization in Authentic Contexts. *Journal of Educational Computing Research*, *61*(5), 1008-1035. https://doi.org/10.1177/07356331221137253

iFlyTek. (n.d.). *Learn about iFLYTEK ETS*. https://edu.iflytek.com/solution/family/ets100

**Jamshed, M., Alam, I., Sultan, S. A., & Banu, S. (2024). Using artificial intelligence for English language learning: Saudi EFL learners' opinions, attitudes and challenges. *Journal of Education and e-Learning Research*, *11*(1), 135-141. https://doi.org/10.20448/jeelr.v11i1.5397

Jeon, J. (2021). Chatbot-assisted dynamic assessment (CA-DA) for L2 vocabulary learning and diagnosis. *Computer Assisted Language Learning*, *36*(7), 1338-1364. https://doi.org/10.1080/09588221.2021.1987272

*Jeon, J. (2023). Chatbot-assisted dynamic assessment (CA-DA) for L2 vocabulary learning and diagnosis. *Computer Assisted Language Learning*, *36*(7), 1338-1364. https://doi.org/10.1080/09588221.2021.1987272

Jiang, K., & Lu, X. (2020). Natural Language Processing and Its Applications in Machine Translation: A Diachronic Review. *2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI)*, 210–214. https://doi.org/10.1109/IICSPI51290.2020.9332458

**Karataş, F., Abedi, F. Y., Gunyel, F. O., Karadeniz, D., & Kuzgun, Y. (2024). Incorporating AI in foreign language education: An investigation into ChatGPT's effect on foreign language learners. *Education and Information Technologies*, 1-24. https://doi.org/10.1007/s10639-024-12574-6

Karnal, A. R., & Pereira, V. V. (2015). Reading Strategies in a L2: A Study on Machine Translation. *The Reading Matrix: An International Online Journal*, *15*(2), 69–79.

**Kemelbekova, Z., Degtyareva, X., Yessenaman, S., Ismailova, D., & Seidaliyeva, G. (2024). AI in teaching English as a foreign language: Effectiveness and prospects in Kazakh higher education. *XLinguae*, *17*(1), 69-83. https://doi.org/10.18355/XL.2024.17.01.05

*Khampusaen, D., Chanprasopchai, T., & Lao-Un, J. (2023). Empowering Thai Community-based Tourism Operators: Enhancing English Pronunciation Abilities with AI-based Lessons. *Journal of Mekong Societies*, *19*(1), 132-159.

**Kharis, M., Schön, S., Hidayat, E., Ardiansyah, R., & Ebner, M. (2022). Mobile Gramabot: Development of a Chatbot App for Interactive German Grammar Learning. *International Journal of Emerging Technologies in Learning*, *17*(14), 52-63. https://doi.org/10.3991/ijet.v17i14.31323

**Kim, A., & Su, Y. (2024). How implementing an AI chatbot impacts Korean as a foreign language learners' willingness to communicate in Korean. *System*, *122*, 1-12. https://doi.org/10.1016/j.system.2024.103256

**Kim, H.-S., & Cha, Y. (2023). The Role of AI Translators on Reading Comprehension. *Korean Journal of English Language and Linguistics*, *23*, 38-58. https://doi.org/10.15738/kjell.23..202301.38

**Kim, H.-S., Cha, Y., & Kim, N. Y. (2021a). Effects of AI chatbots on EFL students' communication skills. *Korean Journal of English Language and Linguistics*, *2021*(21), 712-734. https://doi.org/10.15738/kjell.21..202108.712

**Kim, H.-S., Kim, N., & Cha, Y. (2021b). Is It Beneficial to Use AI Chatbots to Improve Learners' Speaking Performance? *Journal of Asia TEFL*, *18*(1), 161-178. https://doi.org/10.18823/asiatefl.2021.18.1.10.161

**Kim, H.-S., Kim, N.-Y., & Cha, Y. (2022). A Study on the Use of AI-based Learning Programs by EFL Students with Different Types of Teacher Support. *Korean Journal of English Language and Linguistics*, *22*, 355-376. https://doi.org/10.15738/kjell.22..202204.355

Kim, N.-Y. (2018). A study on chatbots for developing Korean college students' English listening and

reading skills. *Journal of Digital Convergence*, *16*(8), 19-26. https://doi.org/10.14400/JDC.2018.16.8.019

Kim, N.-Y. (2019). A Study on the Use of Artificial Intelligence Chatbots for Improving English Grammar Skills. *Journal of Digital Convergence*, *17*(8), 37–46. https://doi.org/10.14400/JDC.2019.17.8.037

**Kim, N.-Y. (2022a). AI-integrated Mobile-assisted Language Learning: Is It an Effective Way of Preparing for the TOEIC Test in Classroom Environments? *English Teaching*, *77*(3), 79-102. https://doi.org/10.15858/engtea.77.3.202209.79

**Kim, N.-Y. (2022b). English with AI: A new era of TOEIC learning for students majoring in airline services. *Linguistic Research*, *39*, 97-122. https://doi.org/10.17250/khisli.39..202209.004

Koehn, P. (2020). *Neural machine translation*. Cambridge University Press.

**Kostikova, I., & Miasoiedova, S. (2019). Supporting post-graduate students writing skills development with the online machine learning tool: Write & Improve. *Information Technologies and Learning Tools*, *74*(6), 238-249.

Krashen, S. D. (1981). *Second language acquisition and second language learning*. Pergamon.

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

Krashen, S. D. (1985). *The input hypothesis*. Longman.

*Kwon, S. K., Shin, D., & Lee, Y. (2023). The application of chatbot as an L2 writing practice tool. *Language Learning & Technology*, *27*(1), 1-19. https://doi.org/10125/73541

Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open*, *6*, 1-13. https://doi.org/10.1016/j.caeo.2024.100174

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. https://doi.org/10.1038/nature14539

*Lee, J. H., Shin, D., & Noh, W. (2023). Artificial Intelligence-Based Content Generator Technology for Young English-as-a-Foreign-Language Learners' Reading Enjoyment. *RELC Journal*, *54*(2), 508-516. https://doi.org/10.1177/00336882231165060

*Lee, S., & Jeon, J. (2022). Visualizing a disembodied agent: young EFL learners' perceptions of voice-controlled conversational agents as language partners. *Computer Assisted Language Learning*, *37* (5-6), 1048-1073. https://doi.org/10.1080/09588221.2022.2067182

Lee, S.-M. (2023). The effectiveness of machine translation in foreign language education: A systematic review and meta-analysis. *Computer Assisted Language Learning*, *36*(1–2), 103–125. https://doi.org/10.1080/09588221.2021.1901745

**Lee, Y.-J., Davis, R. O., & Lee, S. O. (2024). University students' perceptions of artificial intelligence-based tools for English writing courses. *Online Journal of Communication and Media Technologies*, *14*(1), 1-11. https://doi.org/10.30935/ojcmt/14195

**Li, B., & Peng, M. (2021). The Evaluation of a Blended Teaching Mode Based on an AI Language Learning Platform. In ICISE-IE 2021 Organizing Committee (Ed.), *2021 2nd International Conference on Information Science and Education* (pp. 1437-1440). IEEE. https://doi.org/10.1109/ICISE-IE53922.2021.00320

*Li, H., & Graesser, A. C. (2021). The impact of conversational agents' language on summary writing. *Journal of Research on Technology in Education*, *53*(1), 44-66.

https://doi.org/10.1080/15391523.2020.1826022

*Li, X., Li, B., & Cho, S.-J. (2023). Empowering Chinese Language Learners from Low-Income Families to Improve Their Chinese Writing with ChatGPT's Assistance Afterschool. *Languages*, *8*(238), 1-16. https://doi.org/10.3390/languages8040238

Liang, J.-C., Hwang, G.-J., Chen, M.-R. A., & Darmawansah, D. (2023). Roles and research foci of artificial intelligence in language education: An integrated bibliographic analysis and systematic review approach. *Interactive Learning Environments*, *31*(7), 4270–4296. https://doi.org/10.1080/10494820.2021.1958348

**Liao, D. (2023). Reform of English Listening Course Teaching Mode in Vocational Colleges Based on Artificial Intelligence. In T.-H. Meen (Ed.), *2023 IEEE 6th Eurasian Conference on Educational Innovation* (pp. 226-229). IEEE. https://doi.org/10.1109/ECEI57668.2023.10105346

**Lin, C., & Mubarok, H. (2021). Learning Analytics for Investigating the Mind Map-Guided AI Chatbot Approach in an EFL Flipped Speaking Classroom. *Educational Technology & Society*, *24*(4), 16-35.

Link, S., Mehrzad, M., & Rahimi, M. (2022). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning*, *35*(4), 605–634. https://doi.org/10.1080/09588221.2020.1743323

**Liu, C., Hou, J., Tu, Y.-F., Wang, Y., & Hwang, G.-J. (2023). Incorporating a reflective thinking promoting mechanism into artificial intelligence-supported English writing environments. *Interactive Learning Environments*, *31*(9), 5614-5632. https://doi.org/10.1080/10494820.2021.2012812

*Liu, C.-C., Chen, W.-J., Lo, F., Chang, C.-H., & Lin, H.-M. (2024a). Teachable Q&A Agent: The Effect of Chatbot Training by Students on Reading Interest and Engagement. *Journal of Educational Computing Research*, *62*(4), 1122-1154. https://doi.org/10.1177/07356331241236467

*Liu, C.-C., Chiu, C. W., Chang, C.-H., & Lo, F. (2024b). Analysis of a chatbot as a dialogic reading facilitator: its influence on learning interest and learner interactions. *Educational Technology Research and Development*, 1-29. https://doi.org/10.1007/s11423-024-10370-0

*Liu, C.-C., Liao, M.-G., Chang, C.-H., & Lin, H.-M. (2022). An analysis of children' interaction with an AI chatbot and its impact on their interest in reading. *Computers & Education*, *189*, 1-16. https://doi.org/10.1016/j.compedu.2022.104576

*Liu, P.-L., & Chen, C.-J. (2023). Using an AI-Based Object Detection Translation Application for English Vocabulary Learning. *Educational Technology & Society*, *26*(3), 5-20. https://doi.org/10.30191/ETS.202307_26(3).0002

**Liu, X., & Ardakani, S. P. (2022). A machine learning enabled affective E-learning system model. *Education and Information Technologies*, *27*(7), 9913-9934. https://doi.org/10.1007/s10639-022-11010-x

Long, M. H. (1981). Input, interaction and second language acquisition. In H. Winitz (Ed.), *Annals of the New York Academy of Sciences: Vol. XX. Native language and foreign language acquisition* (pp. 259–278). New York Academy of Sciences. https://doi.org/10.1111/j.1749-6632.1981.tb42014.x

Long, M. H. (1985). Input and second language acquisition theory. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 377-393). Newbury House.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of language acquisition* (vol. 2, pp. 413–468). Academic Press.

Luckin, R., Holmes, W., Griffiths, M., & Corcier, L. B. (2016). *Intelligence Unleashed: An argument for AI in Education*. Pearson. http://discovery.ucl.ac.uk/1475756/

Luo, B. (2016). Evaluating a computer-assisted pronunciation training (CAPT) technique for efficient classroom instruction. *Computer Assisted Language Learning*, *29*(3), 451–476. https://doi.org/10.1080/09588221.2014.963123

**Luo, Y., & Liang, W. (2022). Application of POA in Teaching Mode of College English Writing Based on Artificial Intelligence. In T.-H. Meen (Ed.), *5th IEEE Eurasian Conference on Educational Innovation 2022* (pp. 105-109). IEEE. https://doi.org/10.1109/ECEI53102.2022.9829466

Macaro, E. (2019). Systematic reviews in applied linguistics. In J. McKinley & H. Rose (Eds.), *The Routledge Handbook of Research Methods in Applied Linguistics* (pp. 230–239). Routledge. https://doi.org/10.4324/9780367824471-20

Macaro, E., Curle, S., Pun, J., An, J., & Dearden, J. (2018). A systematic review of English medium instruction in higher education. *Language Teaching*, *51*(1), 36–76. https://doi.org/10.1017/S0261444817000350

Macaro, E., Handley, Z., & Walter, C. (2012). A systematic review of CALL in English as a second language: Focus on primary and secondary education. *Language Teaching*, *45*(1), 1–43. https://doi.org/10.1017/S0261444811000395

*Madhavi, E., Sivapurapu, L., Koppula, V., Esther Rani, P. B., & Sreehari, V. (2023). Developing Learners' English Speaking Skills using ICT and AI Tools. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, *32*(2), 142-153. https://doi.org/10.37934/ARASET.32.2.142153

**Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: a mixed methods intervention study. *Smart Learning Environments*, *11*(9), 1-18. https://doi.org/10.1186/s40561-024-00295-9

**Markus, A. M., Ovinova, L. N., Dmitrusenko, I. N., & Shraiber, E. G. (2023). Application of Artificial Intelligence Technology in Teaching English Language to Engineering Bachelors. In S. Shaposhnikov (Ed.), *2023 International Conference on Quality Management, Transport and Information Security, Information Technologies* (pp. 147-151). IEEE. https://doi.org/10.1109/ITQMTIS58985.2023.10346594

McCarthy, J. (2007, November 12). *WHAT IS ARTIFICIAL INTELLIGENCE?* Formal Reasoning Group. https://www-formal.stanford.edu/jmc/whatisai/

McCrocklin, S. M. (2016). Pronunciation learner autonomy: The potential of Automatic Speech Recognition. *System*, *57*, 25–42. https://doi.org/10.1016/j.system.2015.12.013

*Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, *6*, 1-10. https://doi.org/10.1016/j.caeai.2023.100199

*Mohamed, S. S. A., & Alian, E. M. I. (2023). Student's Attitude toward Using Chatbot in EFL Learning. *Arab World English Journal*, *14*(3), 15-27. https://dx.doi.org/10.24093/awej/vol14no3.2

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*(7), 1-6. https://doi.org/10.1371/journal.pmed.1000097

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., & PRISMA-P Group. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, *4*(1), 1-9. https://doi.org/10.1186/2046-4053-4-1

Montalvo, S., Palomo, J., & de la Orden, C. (2018). Building an educational platform using NLP: a case study in teaching finance. *Journal of Universal Computer Science, 24*(10), 1403-1423.

Mousavinasab, E., Zarifsanaiey, N., R. Niakan Kalhori, S., Rakhshan, M., Keikha, L., & Ghazi Saeedi, M. (2021). Intelligent tutoring systems: A systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments*, *29*(1), 142–163. https://doi.org/10.1080/10494820.2018.1558257

**Muniandy, J., & Selvanathan, M. (2024). ChatGPT, a partnering tool to improve ESL learners' speaking skills: Case study in a Public University, Malaysia. *Teaching Public Administration*, 1-17. https://doi.org/10.1177/01447394241230152

**Nagata, N. (1995). An effective application of natural language processing in second language instruction. *CALICO Journal*, *13*(1), 47-67. https://doi.org/10.1558/cj.v13i1.47-67

**Nghi, T. T., Phuc, T. H., & Thang, N. T. (2019). Applying AI chatbot for teaching a foreign language: An empirical research. *International Journal of Scientific and Technology Research*, *8*(12), 897-902.

Niño, A. (2008). Evaluating the use of machine translation post-editing in the foreign language class. *Computer Assisted Language Learning*, *21*(1), 29–49. https://doi.org/10.1080/09588220701865482

Niño, A. (2009). Machine translation in foreign language learning: Language learners' and tutors' perceptions of its advantages and disadvantages. *ReCALL*, *21*(2), 241–258. https://doi.org/10.1017/S0958344009000172

*Nong, L., Liu, G., & Tan, C. (2021). An Empirical Study on the Implementation of AI Assisted Language Teaching for Improving Learner's Learning Ability. In H. Zhan, G. Liu, & J. Hall (Eds.), *2021 Tenth International Conference of Educational Innovation through Technology* (pp. 215-221). IEEE. https://doi.org/10.1109/EITT53287.2021.00050

**Nozhovnik, O., Harbuza, T., Teslenko, N., Okhrimenko, O., Zalizniuk, V., & Durdas, A. (2023). Chatbot Gamified and Automated Management of L2 Learning Process Using Smart Sender Platform. *International Journal of Educational Methodology*, *9*(3), 603-618. https://doi.org/10.12973/ijem.9.3.603

OpenAI. (2015, December 11). *Introducing OpenAI*. https://openai.com/index/introducing-openai/

OpenAI. (2022, November 30). *Introducing ChatGPT*. https://openai.com/index/chatgpt/

OpenAI. (2024, May 13). Introducing GPT-4o and more tools to ChatGPT free users. https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/

**Oraif, I. (2024). Natural Language Processing (NLP) and EFL Learning: A Case Study Based on Deep Learning. *Journal of Language Teaching and Research*, *15*(1), 201-208. https://doi.org/10.17507/jltr.1501.22

Ouyang, F., Wu, M., Zheng, L., Zhang, L., & Jiao, P. (2023). Integration of artificial intelligence performance prediction and learning analytics to improve student learning in online engineering

course. *International Journal of Educational Technology in Higher Education*, *20*(1), 1-23. https://doi.org/10.1186/s41239-022-00372-4

Ouyang, F., Zheng, L., & Jiao, P. (2022). Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020. *Education and Information Technologies*, *27*(6), 7893–7925. https://doi.org/10.1007/s10639-022-10925-9

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan -- a web and mobile app for systematic reviews. *Systematic Reviews*, 5(1), 1-10. https://doi.org/10.1186/s13643-016-0384-4

Pace, R., Pluye, P., Bartlett, G., Macaulay, A. C., Salsberg, J., Jagosh, J., & Seller, R. (2012). Testing the reliability and efficiency of the pilot Mixed Methods Appraisal Tool (MMAT) for systematic mixed studies review. *International Journal of Nursing Studies*, *49*(1), 47–53. https://doi.org/10.1016/j.ijnurstu.2011.07.002

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., … Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ (Clinical research ed.)*, *372*, 1-9. https://doi.org/10.1136/bmj.n71

**Peña-Acuña, B., & Crismán-Pérez, R. (2022). Research on Papua, a digital tool with artificial intelligence in favor of learning and linguistic attitudes towards the learning of the English language in students of Spanish language as L1. *Frontiers in Psychology*, *13*, 1-15. https://doi.org/10.3389/fpsyg.2022.1019278

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Blackwell Publishing. https://doi.org/10.1002/9780470754887

Plotly Technologies Inc. (2015). *Collaborative data science*. Montréal, QC. https://plot.ly

Popenici, S. A. D., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, *12*(22), 1-13. https://doi.org/10.1186/s41039-017-0062-8

**Qasem, F., Ghaleb, M., Mahdi, H. S., Khateeb, A. A., & Fadda, H. A. (2023). Dialog chatbot as an interactive online tool in enhancing ESP vocabulary learning. *Saudi Journal of Language Studies*, *3*(2), 76-86. https://doi.org/10.1108/SJLS-10-2022-0072

Qian, K., Shea, R., Li, Y., Fryer, L.K., & Yu, Z. (2023). User Adaptive Language Learning Chatbots with a Curriculum. In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, & O. C. Santos (Eds.), *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky* (pp. 308–313). Springer. https://doi.org/10.1007/978-3-031-36336-8_48

**Qiao, H., & Zhao, A. (2023). Artificial intelligence-based language learning: Illuminating the impact on speaking skills and self-regulation in Chinese EFL context. *Frontiers in Psychology*, *14*, 1–15. https://doi.org/10.3389/fpsyg.2023.1255594

Roscoe, R. D., & McNamara, D. S. (2013). Writing Pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology*, 105(4), 1010–1025. https://doi.org/10.1037/a0032340

Russell, S. J., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd ed.). Prentice Hall.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of*

Research and Development, *3*(3), 210–229. https://doi.org/10.1147/rd.33.0210

Sarker, I. H. (2021a). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, *2*(420), 1-20. https://doi.org/10.1007/s42979-021-00815-1

Sarker, I. H. (2021b). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(160), 1-21. https://doi.org/10.1007/s42979-021-00592-x

Schulz, J., Hamilton, C., Wonnacott, E., & Murphy, V. (2023). The impact of multi-word units in early foreign language learning and teaching contexts: A systematic review. *Review of Education*, *11*, 1-26. https://doi.org/10.1002/rev3.3413

Shadiev, R., & Liu, J. (2023). Review of research on applications of speech recognition technology to assist language learning. *ReCALL*, *35*(1), 74–88. https://doi.org/10.1017/S095834402200012X

**Shafiee Rad, H. (2024). Revolutionizing L2 speaking proficiency, willingness to communicate, and perceptions through artificial intelligence: A case of Speeko application. *Innovation in Language Learning and Teaching*, 1–16. https://doi.org/10.1080/17501229.2024.2309539

*Shafiee Rad, H., & Roohani, A. (2024). Fostering L2 Learners' Pronunciation and Motivation via Affordances of Artificial Intelligence. *Computers in the Schools*, 1-23. https://doi.org/10.1080/07380569.2024.2330427

**Shafiee Rad, H., Alipour, R., & Jafarpour, A. (2023). Using artificial intelligence to foster students' writing feedback literacy, engagement, and outcome: A case of Wordtune application. *Interactive Learning Environments*, 1–21. https://doi.org/10.1080/10494820.2023.2208170

**Shaikh, S., Yayilgan, S. Y., Klimova, B., & Pikhart, M. (2023). Assessing the Usability of ChatGPT for Formal English Language Learning. *European Journal of Investigation in Health, Psychology and Education*, *13*(9), 1937-1960. https://doi.org/10.3390/ejihpe13090140

Shermis, M. D., Burstein, J., & Bursky, S. A. (2013). Introduction to automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 1–15). Routledge.

Shute, V. J., & Psotka, J. (1996). Intelligent Tutoring System: Past, Present, and Future. In D. Jonassen (Ed.), *Handbook of research for educational communications and technology* (pp. 570–600). Macmillan.

Skehan, P. (1998). *A Cognitive Approach to Language Learning*. Oxford University Press.

Son, J.-B., Ružić, N. K., & Philpott, A. (2023). Artificial intelligence technologies and applications for language learning and teaching. *Journal of China Computer-Assisted Language Learning*, 1-19. https://doi.org/10.1515/jccall-2023-0015

**Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology*, *14*, 1-14. https://doi.org/10.3389/fpsyg.2023.1260843

*Srinivasan, V., & Murthy, H. (2021). Improving reading and comprehension in K-12: Evidence from a large-scale AI technology intervention in India. *Computers and Education: Artificial Intelligence*, *2*, 1-22. https://doi.org/10.1016/j.caeai.2021.100019

Stewart, L., Moher, D., & Shekelle, P. (2012). Why prospective registration of systematic reviews makes sense. *Systematic Reviews*, *1*(7), 1–4. https://doi.org/10.1186/2046-4053-1-7

**Sun, L. (2022). The Application of Computer Aid in English Teaching under the Background of

Artificial Intelligence-Based on Awrite Survey Data Analysis. In C. Ceballos (Ed.), *2022 International Conference on Education, Network and Information Technology* (pp. 160-164). IEEE. https://doi.org/10.1109/ICENIT57306.2022.00042

Sze, V., Chen, Y.-H., Yang, T.-J., & Emer, J. S. (2017). Efficient Processing of Deep Neural Networks: A Tutorial and Survey. *Proceedings of the IEEE*, *105*(12), 2295–2329. https://doi.org/10.1109/JPROC.2017.2761740

**Tegos, S., Demetriadis, S., & Tsiatsos, T. (2014). A configurable conversational agent to trigger students' productive dialogue: A pilot study in the CALL domain. *International Journal of Artificial Intelligence in Education*, *24*(1), 62-91. https://doi.org/10.1007/s40593-013-0007-3

Teo, A. (2012). Promoting EFL students' inferential reading skills through computerised dynamic assessment. *Language Learning & Technology*, *16*(3), 10-20. http://dx.doi.org/10125/44292

Tsai, S.-C. (2019). Using google translate in EFL drafts: A preliminary investigation. *Computer Assisted Language Learning*, *32*(5–6), 510–526. https://doi.org/10.1080/09588221.2018.1527361

UNESCO. (2021). *AI and education: guidance for policy-makers*. https://unesdoc.unesco.org/ark:/48223/pf0000376709

UNESCO. (2022a). *K-12 AI curricula: a mapping of government-endorsed AI curricula*. https://unesdoc.unesco.org/ark:/48223/pf0000380602

UNESCO. (2022b). International forum on AI and education: steering AI to empower teachers and transform teaching, 5-6 December 2022; analytical report. https://unesdoc.unesco.org/ark:/48223/pf0000386162

UNESCO. (2023). *Guidance for generative AI in education and research*. https://unesdoc.unesco.org/ark:/48223/pf0000386693

Valijärvi, R.-L., & Tarsoly, E. (2019). Translating Google Translate to the language classroom: pitfalls and possibilities. *Practitioner Research in Higher Education*, *12* (1), 61–74.

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523-538. https://doi.org/10.1007/s11192-009-0146-3

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (arXiv:1706.03762). arXiv. http://arxiv.org/abs/1706.03762

**Vu, D. C., Lian, A.-P., & Siriyothin, P. (2022). Integrating Natural Language Processing and Multimedia Databases in CALL Software: Development and Evaluation of an ICALL Application for EFL Listening Comprehension. *Computer Assisted Language Learning Electronic Journal (CALL-EJ)*, *23*(3), 41-69.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological process*. Harvard University Press.

Vygotsky, L. S. (1986). *Thought and language*. MIT Press.

**Wale, B. D. (2024). Artificial intelligence in education: Effects of using integrative automated writing evaluation programs on honing academic writing instruction. *Cakrawala Pendidikan*, *43*(1), 273-287. https://doi.org/10.21831/cp.v43i1.67715

**Wale, B., & Kassahun, Y. (2024). The Transformative Power of AI Writing Technologies: Enhancing EFL Writing Instruction through the Integrative Use of Writerly and Google Docs. *Human Behavior and Emerging Technologies*, *2024*, 1-15.

https://doi.org/10.1155/2024/9221377

Wang, P. (2019). On Defining Artificial Intelligence. *Journal of Artificial General Intelligence*, *10*(2), 1–37. https://doi.org/10.2478/jagi-2019-0002

Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T., & Du, Z. (2024). Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications*, *252*, 1-19. https://doi.org/10.1016/j.eswa.2024.124167

**Wang, X., & Feng, Y. (2023). An Experimental Study of ChatGPT-Assisted Improvement of Chinese College Students' English Reading Skills: A Case Study of Dear Life. In 15th International Conference on Education Technology and Computers (pp. 21-26). ACM. https://doi.org/10.1145/3629296.3629300

**Wang, Z. (2022). Computer-assisted EFL writing and evaluations based on artificial intelligence: a case from a college reading and writing course. *Library Hi Tech*, *40*(1), 80-97. https://doi.org/10.1108/LHT-05-2020-0113

Webb, M. E., Fluck, A., Magenheim, J., Malyn-Smith, J., Waters, J., Deschênes, M., & Zagami, J. (2021). Machine learning for human learners: Opportunities, issues, tensions, and threats. *Educational Technology Research and Development*, *69*, 2109–2130. https://doi.org/10.1007/s11423-020-09858-2

**Wei, L. (2023). Artificial intelligence in language instruction: Impact on English learning achievement, L2 motivation, and self-regulated learning. *Frontiers in Psychology*, *14*, 1–14. https://doi.org/10.3389/fpsyg.2023.1261955

Weizenbaum, J. (1966). ELIZA -- a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36-45. https://doi.org/10.1145/365153. 365168

**Wiboolyasarin, W., Wiboolyasarin, K., Suwanwihok, K., Jinowat, N., & Muenjanchoey, R. (2024). Synergizing collaborative writing and AI feedback: An investigation into enhancing L2 writing proficiency in wiki-based environments. *Computers and Education: Artificial Intelligence*, *6*, 1-10. https://doi.org/10.1016/j.caeai.2024.100228

Willis, S., Neil, R., Mellick, M. C., & Wasley, D. (2019). The Relationship Between Occupational Demands and Well-Being of Performing Artists: A Systematic Review. *Frontiers in Psychology*, *10*, 1-20. https://doi.org/10.3389/fpsyg.2019.00393

**Wu, L. (2022). Case Study on Application of Artificial Intelligence to Oral English Teaching in Vocational Colleges. In T.-H. Meen (Ed.), *2022 International Conference on Computation, Big-Data and Engineering* (pp. 71-74). IEEE. https://doi.org/10.1109/ICCBE56101.2022.9888213

Wu, W.-C. V., Hsieh, J. S. C., & Yang, J. C. (2017). Creating an online learning community in a flipped classroom to enhance EFL learners' oral proficiency. *Educational Technology & Society*, *20*(2), 142–157.

**Xiao, Y., & Zhi, Y. (2023). An Exploratory Study of EFL Learners' Use of ChatGPT for Language Learning Tasks: Experience and Perceptions. *Languages*, *8*(3), 1-12. https://doi.org/10.3390/languages8030212

Xie, H., Chu, H.-C., Hwang, G.-J., & Wang, C.-C. (2019). Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of journal publications from 2007 to 2017. *Computers & Education*, *140*, 1-16. https://doi.org/10.1016/j.compedu.2019.103599

Xing, L.-S. (2023). Artificial Intelligence and Digitalization in China's Education System: A Systematic Analysis of the Policy Framework and Underlying Strategies. *Working Papers on*

East Asian Studies, *136*, pp. 1-48. https://doi.org/10.17185/DUEPUBLICO/78369

Xu, Z., Wijekumar, K. (Kay), Ramirez, G., Hu, X., & Irey, R. (2019). The effectiveness of intelligent tutoring systems on K-12 students' reading comprehension: A meta-analysis. *British Journal of Educational Technology*, *50*(6), 3119–3137. https://doi.org/10.1111/bjet.12758

Yang, H., & Kyun, S. (2022). The current research trend of artificial intelligence in language learning: A systematic empirical literature review from an activity theory perspective. *Australasian Journal of Educational Technology*, *38*(5), 180–210. https://doi.org/10.14742/ajet.7492

*Yang, H., Kim, H., Lee, J. H., & Shin, D. (2022). Implementation of an AI chatbot as an English conversation partner in EFL speaking classes. *ReCALL*, *34*(3), 327-343. https://doi.org/10.1017/S0958344022000039

*Ye, Y., Deng, J., Liang, Q., & Liu, X. (2022). Using a Smartphone-Based Chatbot in EFL Learners' Oral Tasks. *International Journal of Mobile and Blended Learning*, *14*(1), 1-17. https://doi.org/10.4018/IJMBL.299405

Yenduri, G., Ramalingam, M., Selvi, G. C., Supriya, Y., Srivastava, G., Maddikunta, P. K. R., Raj, G. D., Jhaveri, R. H., Prabadevi, B., Wang, W., Vasilakos, A. V., & Gadekallu, T. R. (2024). GPT (Generative Pre-Trained Transformer)—A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. *IEEE Access*, *12*, 54608–54649. https://doi.org/10.1109/ACCESS.2024.3389497

Young, V., & Mihailidis, A. (2010). Difficulties in Automatic Speech Recognition of Dysarthric Speakers and Implications for Speech-Based Applications Used by the Elderly: A Literature Review. *Assistive Technology*, *22*(2), 99–112. https://doi.org/10.1080/10400435.2010.483646

*Yuan, Y. (2023). An empirical study of the efficacy of AI chatbots for English as a foreign language learning in primary education. *Interactive Learning Environments*, 1–16. https://doi.org/10.1080/10494820.2023.2282112

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(39), 1-27. https://doi.org/10.1186/s41239-019-0171-0

**Zhang, C., Meng, Y., & Ma, X. (2024). Artificial intelligence in EFL speaking: Impact on enjoyment, anxiety, and willingness to communicate. *System*, *121*, 1-14. https://doi.org/10.1016/j.system.2024.103259

Zhang, C., Yu, T., Gao, Y., & Tham, M. L. (2023). Design of a Smart Teaching English Translation System Based on Big Data Machine Learning. *International Journal of Web-Based Learning and Teaching Technologies*, *18*(2), 1–14. https://doi.org/10.4018/IJWLTT.330144

Zhang, R., Zou, D., & Cheng, G. (2023a). A review of chatbot-assisted learning: Pedagogical approaches, implementations, factors leading to effectiveness, theories, and future directions. *Interactive Learning Environments*, 1–29. https://doi.org/10.1080/10494820.2023.2202704

**Zhang, R., Zou, D., & Cheng, G. (2023b). Chatbot-based learning of logical fallacies in EFL writing: perceived effectiveness in improving target knowledge and learner motivation. *Interactive Learning Environments*, 1-18. https://doi.org/10.1080/10494820.2023.2220374

**Zhang, R., Zou, D., & Cheng, G. (2023c). Chatbot-based training on logical fallacy in EFL argumentative writing. *Innovation in Language Learning and Teaching*, *17*(5), 932-945. https://doi.org/10.1080/17501229.2023.2197417

Zhang, W., Cai, M., Lee, H. J., Evans, R., Zhu, C., & Ming, C. (2024). AI in Medical Education: Global situation, effects and challenges. *Education and Information Technologies*, *29*, 4611–4633. https://doi.org/10.1007/s10639-023-12009-8

**Zhang, Y. (2024). A lesson study on a MOOC-based and AI-powered flipped teaching and assessment of EFL writing model: teachers' and students' growth. *International Journal for Lesson and Learning Studies*, *13*(1), 28-40. https://doi.org/10.1108/IJLLS-07-2023-0085

*Zhang, Z., & Huang, X. (2024). The impact of chatbots based on large language models on second language vocabulary acquisition. *Heliyon*, *10*(3), 1-13. https://doi.org/10.1016/j.heliyon.2024.e25370

Zhao, R., Zhuang, Y., Zou, D., Xie, Q., & Yu, P. L. H. (2023). AI-assisted automated scoring of picture-cued writing tasks for language assessment. *Education and Information Technologies*, *28*, 7031–7063. https://doi.org/10.1007/s10639-022-11473-y

Zheng, L., Niu, J., Zhong, L., & Gyasi, J. F. (2023). The effectiveness of artificial intelligence on learning achievement and learning perception: A meta-analysis. *Interactive Learning Environments*, *31*(9), 5650-5664. https://doi.org/10.1080/10494820.2021.2015693

**Zhou, C. (2023). Integration of modern technologies in higher education on the example of artificial intelligence use. *Education and Information Technologies*, *28*(4), 3893–3910. https://doi.org/10.1007/s10639-022-11309-9

**Zou, B., Guan, X., Shao, Y., & Chen, P. (2023). Supporting Speaking Practice by Social Network-Based Interaction in Artificial Intelligence (AI)-Assisted Language Learning. *Sustainability*, *15*(4). https://doi.org/10.3390/su15042872

Appendix 1. The review protocol

Title

The effectiveness of using artificial intelligence-assisted tools in second and foreign language learning: a systematic review

1 Introduction

1.1 Rationale

Artificial Intelligence (AI) technologies have tremendously influenced the educational sector (Luckin et al., 2016; UNESCO, 2021; UNESCO, 2023). The rapid advancement in AI techniques, such as machine learning, deep learning, and natural language processing, has provided momentum for the application of AI in language education (Zawacki-Richter et al., 2019). As a result, AI technologies and their applications, such as chatbots, intelligent tutoring systems, and automated scoring systems, have been increasingly adopted to support language learning and teaching (Chen et al., 2020). With the launch of ChatGPT, an AI chatbot powered by large language models (LLMs), in late 2022, AI-assisted language education has garnered ever-greater scholarly attention (An et al., 2023). Empirical studies have shown the benefits of AI-supported tools for language learners, including improving learning outcomes (Fryer et al., 2020), providing abundant learning resources (Zou et al., 2020), and enabling personalised learning (Wang et al., 2021).

Despite numerous findings on the effectiveness of AI-powered language learning applications, systematic examinations on the topic, particularly in language education rather than general education, remain scarce. Another research gap is that few previous reviews synthesised studies on LLMs, which might be insufficient for understanding the latest trends of emerging AI technologies and their applications. To address these limitations, this review aims to include studies dating from 1956, when the concept of AI was proposed, to more recent studies on generative AI. The review will synthesize studies on how second and foreign language learners use AI-driven tools and applications to assist their language learning. It is hoped that the current review could provide more comprehensive insights into the research field.

1.2 Objective

This study aims to systematically review publications on the effectiveness of AI-assisted second and foreign language learning. The review will collect studies on the impact of AI-based learning tools (e.g., chatbots, intelligent tutoring systems, etc.) on second or foreign language learners' learning outcomes and learning perceptions. In these studies, learning outcomes are measured by standardised

tests, and learning perceptions encompass learners' attitudes, motivation, and opinions on the learning experience (Zheng et al., 2021). Specifically, the following research questions will be addressed:

RQ1. What is the extent of research that has been conducted to investigate the effectiveness of AI-assisted tools in second and foreign language learning?

1a. What AI technologies and their applications have been researched?

1b. What language skills in AI-supported language learning have been researched?

RQ2[3]. What is the extent of the evidence of the effectiveness of AI-assisted second and foreign language learning?

2a. What are the effects of AI-assisted learning tools on language learners' learning outcomes?

2b. What are the effects of AI-assisted learning tools on language learners' learning perceptions?

2 Method

2.1 Eligibility criteria

To address the review questions, the researcher defined the inclusion and exclusion criteria for the current review based on PICOSS elements[4] (Cherry & Dickson, 2017). Table 1 provides the eligibility criteria and rationale.

Table 1. Eligibility criteria for this review

Criteria	Inclusion	Exclusion	Rationale
Reference	Studies with complete bibliographic information	Studies with incomplete bibliographic information	The researcher needs a complete reference to retrieve the studies that are reviewed.
Time period	Studies published in or after 1956	Studies published before 1956	The concept of artificial intelligence was originally proposed in 1956 by John McCarthy.

[3] If a further meta-analysis could be conducted, a third sub-question could be how the effectiveness of AI-assisted language learning tools is moderated by different variables.

[4] PICOSS elements include population, intervention, comparator, outcomes, study setting, and setting. More criteria elements (e.g., language of publication, type of publication, etc.) will be considered in this review.

Language of publication	Publications in English and Chinese	Publications in languages other than English and Chinese	The researcher is fully aware of the language bias caused by only including publications in certain languages. However, the researcher could only read English and Chinese.
Type of publication	Peer-reviewed literature and doctoral dissertations	Other publications that are not peer-reviewed (i.e., grey literature)	Peer-reviewed publications have been revised to achieve the best possible validity and accuracy before publication. Besides, Macaro (2019) suggested including doctoral theses because they have undergone rigorous scrutiny by examiners.
Participants	Studies on typically developing second and foreign language learners of any age, gender, L1 background, and second/foreign language proficiency level	1) Studies on non-typically developing learners, such as those with Developmental Language Disorder; 2) Studies on typically developing L1 learners	Findings on non-typically developing learners may not be generalisable to the larger population. Additionally, preliminary searches have yielded few studies on using AI-assisted tools for L1 acquisition; thus, this review focuses on the effect of AI-powered

			applications on second and foreign language learning.
Intervention	Studies that used all types of AI-supported learning tools (e.g., chatbots, automated tutors, intelligent tutoring systems, language learning applications equipped with AI technologies, etc.) to assist second/ foreign language learning	1) Studies that adopted AI-based learning tools to develop learners' first language (L1) skills, such as literacy, and other skills; 2) Studies that used AI-based tools as an intervention to assist language teaching activities (e.g., assessment, feedback provision) or classroom management	The review does not focus on the effect of AI-assisted learning tools on L1 learners' learning achievements. Moreover, the review does not focus on language teachers' use or perceptions of these applications.
Outcomes	Studies reporting empirical data on the effectiveness of AI-assisted applications in second and foreign language learning (e.g., learning outcomes measured by standardised tests, learners' attitudes collected by questionnaires and interviews)	Studies that did not report empirical data collected from second and foreign language learners	The comparison of substantive learning outcomes (e.g., test scores, questionnaire data) could address the research question about the effectiveness of AI-assisted tools in second and foreign language learning

Study design	1) Randomized controlled trial (RCT) and non-RCT in which the experimental group used AI to assist second/foreign language learning, while the control group did not use AI to assist second/foreign language learning; 2) Cohort study in which one group of participants adopting AI-assisted language learning was pre-and-post tested and (/or) post-interviewed 3) Cohort study in which one group of participants adopting AI-assisted language learning was only post-tested and (/or) post-interviewed 4) Cross-sectional survey study in which data were collected at one point in time; 5) Mixed methods studies that combined the abovementioned study designs	1) Studies on the AI algorithms of learning applications that do not involve human participants; 2) Systematic reviews and meta-analyses	Experimental and quasi-experimental studies are most reliable for answering research questions on the effectiveness of AI-assisted language learning tools. Survey studies were also included because they were found relevant in preliminary scoping searches related to learning perceptions, although they may not add to findings on the effectiveness of these applications.
Setting	1) Studies in laboratory or classroom settings where second/foreign language learners use AI-assisted tools;	Studies in laboratory or classroom settings where AI-assisted tools are used to address teachers' needs.	Laboratory and in- and out-of-classroom settings are included because the reviewed publications should be

	2) Studies in which second/foreign language learners use AI-assisted tools for self-study in various out-of-class settings		as comprehensive as possible.

2.2 Information sources

To identify the most relevant databases for the current review, the researcher booked a literature search consultation with an experienced librarian at Bodleian Education Library on November 10, 2023. After the consultation, the researcher generated a list of databases covering the following subjects: (1) education, (2) linguistics, (3) psychology, and (4) multidiscipline (see Table 2). Dissertation databases were also added to the list because Macro (2019) suggested including doctoral theses in systematic reviews. All databases will be accessed electronically through Search Oxford Libraries Online (SOLO), the online catalogue for library collections at the University of Oxford. In addition to literature searches in electronic databases, the researcher will conduct forward and backward citation chaining to complement the search results. Once identifying a paper that meets inclusion criteria, the researcher will conduct a forward search for eligible studies among subsequent publications that have cited the paper. The researcher will also examine the paper's reference list to look for relevant studies in backward searching (i.e. 'snowballing'). The researcher will update the search results every two weeks. The last search is expected to be conducted in the final week of May 2024.

Table 2. A list of consulted databases

Category	**Database**
Education	British Education Index
	Education Collection (incl. ERIC)
Linguistics	Linguistics Collection (including LLBA)
Psychology	PsycINFO
Multidiscipline	SCOPUS
	Web of Science Core Collection

Dissertation	ProQuest Dissertations and Theses
	Oxford University Research Archive (ORA)
	China Academic Journals (CNKI) - Chinese

2.3 Search strategy

Search items were generated based on previous systematic reviews on AI in education (e.g., Ji et al., 2023; Liang et al., 2023; Zheng et al., 2021). At the initial stage, the following keywords were identified for pilot searches: (1) words related to AI: "artificial intelligence" OR AI OR "artificial neural network" OR "machine intelligence" OR "machine learning" OR "deep learning" OR robotic OR "intelligent support" OR "intelligent virtual reality" OR chatbot OR "chat bot" OR "conversational agent" OR "automated tutor" OR "personal tutor" OR "intelligent agent" OR "intelligent system" OR "intelligent tutor" OR "natural language processing"; (2) words related to language learning: language OR "second language" OR L2 OR "foreign language" OR pronunciation OR vocabulary OR grammar OR listening OR speaking OR reading OR writing; (3) words related to effect: outcome* OR achieve* OR perform* OR effect* OR impact*. The asterisk is used as a truncation symbol to search for free-text words with the same root but alternative endings. For example, a keyword search for *achieve** will retrieve search results containing *achieve*, *achieves*, *achieved*, *achievement*, and *achievements*.

After piloting the search, two modifications were made to the search items. Firstly, given the increasing popularity of generative AI[5] and large language models, the first cluster of search terms relevant to AI included "generative artificial intelligence" OR "generative AI" OR GenAI OR "large language model" OR LLM OR ChatGPT OR GPT-3.5 OR GPT-4. Secondly, to ensure a manageable amount of search output, "language" in the second group of search items was refined to "language learn*" OR "language education" according to the librarian's suggestion, because "language" easily appeared in irrelevant publications. The final search terms are shown in Table 3.

Table 3. Search words

Key terms		
Artificial intelligence	**Language learning**	**Effect**
artificial intelligence	language learn*	outcome*
AI	language education	achieve*
artificial neural network	second language	perform*

[5] Generative AI refers to the AI systems that function to generate content.

machine intelligence	L2	effect*
machine learning	foreign language	impact*
deep learning	pronunciation	
robotic	vocabulary	
intelligent support	grammar	
intelligent virtual reality	listening	
chatbot	speaking	
chat bot	reading	
conversational agent	writing	
automated tutor		
personal tutor		
intelligent agent		
intelligent system		
intelligent tutor		
natural language processing		
generative artificial intelligence		
generative AI		
GenAI		
large language model		
LLM		
ChatGPT		
GPT-3.5		
GPT-4		

To avoid unmanageable search results, words related to AI were assigned to the search frame of TITLE, while words related to language learning and effect were assigned to FULL TEXT. The author decided on the search frames with the librarian after more than five search trials. Boolean operators “AND” and “OR” were used to connect the search items.

2.4 Data management

Three tools will be used to manage the search data: (1) Rayyan, (2) EndNote, and (3) Microsoft Excel. How each tool is used will be specified in the next section.

2.5 Data screening and selection process

2.5.1 Title and abstract screening

All search results will be exported as RIS files and uploaded to Rayyan (Ouzzani et al., 2016), a web tool for data screening and management in systematic reviews. Firstly, the duplicate items will be deleted. Secondly, the researcher will examine the titles and abstracts of the collected articles to decide whether to include them or not according to eligibility criteria (cf. Table 1). Articles that meet inclusion criteria will be marked as "include". Articles that are yet to be decided on inclusion or exclusion will be tagged as "include". Articles that fail to meet inclusion criteria will be marked as "exclude". Reasons for exclusion will be specified, including no empirical data, wrong study design, etc.

2.5.2 Full-text screening

The articles labelled as "included" will then be retrieved for their full texts through SOLO and academic platforms such as ResearchGate. All full texts will then be exported to Endnote, a reference management tool, for screening. A Microsoft Excel spreadsheet will be used to record the decision for inclusion or exclusion as well as reasons for the decision.

2.5.3 Citation chaining

As mentioned in Section 2.2, the researcher will conduct forward and backward searches for potentially eligible articles. These texts will also be screened according to the inclusion and exclusion criteria.

2.5.4 Quality assessment

To ensure the reliability and validity of eligibility criteria, a second reviewer who studies MSc Applied Linguistics and Second Language Acquisition at the University of Oxford will be invited to participate in both phases of screening. The second reviewer will review 10% of the total search results. Her decision should be compared with the main researcher's decision to see if 90% agreement can be achieved.

To demonstrate the screening and selection process, a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram (Moher et al., 2009) will be presented.

2.6 Data extraction

The researcher has designed a data extraction form on Microsoft Excel to record relevant data that may answer the research questions. Following Fleeman and Dundar's (2017) guide, the form will be piloted among at least three included studies to check whether it is effective in extracting data. The second reviewer will also independently skim 10% of the included studies. The data extracted by the

two reviewers will be compared to check if there exists any disagreement. The disagreements will then be resolved through discussion.

2.7 Risk of bias assessment

Quality assessment of individual studies will be conducted after data extraction to reduce reporting bias. The 2018 version of Mixed Methods Appraisal Tool (MMAT; Hong et al., 2018) will be used to assess the methodological quality of the included studies. The quality assessment tool was chosen because it is suitable for appraising quantitative, qualitative, and mixed methods studies with different research designs, which facilitates standardised assessment across studies (Pace et al., 2012). Additionally, this appraisal tool has been used in previous systematic reviews in various topic areas (e.g., Schulz et al., 2023; Willis et al., 2019). Specifically, MMAT can be used with five categories of studies: (1) qualitative studies, (2) quantitative randomised controlled trials, (3) quantitative non-randomised studies, (4) quantitative descriptive studies, and (5) mixed methods studies. Five quality criteria are provided for assessing each of the five categories, and each criterion can be marked "yes", "no", or "can't tell".

Following Hong et al. (2020), an overall quality score for each study can be presented using stars (*) or percentages (%). For a quantitative or qualitative study, the score can be:

5***** or 100% quality criteria met

4 **** or 80% quality criteria met

3 *** or 60% quality criteria met

2 ** or 40% quality criteria met

1 * or 20% quality criteria met

For a mixed methods study which needs to be assessed by 15 criteria, its overall quality score equates to the lowest score of the study's components. Therefore, the score is:

20% (*) when QUAL[6]=1 or QUAN=1 or MM=1

40% (**) when QUAL=2 or QUAN=2 or MM=2

60% (***) when QUAL=3 or QUAN=3 or MM=3

80% (****) when QUAL=4 and QUAN=4 and MM=4

100% (*****) when QUAL=5 or QUAN=5 or MM=5

The researcher is aware that such quantification provides an uninformative picture of which parts of studies are problematic, but scoring makes it easier to report the quality assessment results. The second reviewer will assess 10% of the studies.

[6] QUAL represents the score of the qualitative category. QUAN represents the score of the quantitative category. MM represents the score of the mixed methods category.

2.8 Data synthesis

The synthesis will follow Petticrew and Roberts' (2006) framework to present a narrative summary of the data. The three steps include (1) grouping studies into logical categories; (2) analysing findings within each category; and (3) summarising findings across studies. The included studies will also be checked for their homogeneity to decide whether it is appropriate to combine the studies in a meta-analysis. If so, SPSS will be used for statistical synthesis.

2.9 Confidence in cumulative evidence

The overall quality of evidence (also referred to as "the certainty of evidence" or "the confidence in the effect estimates") will be evaluated with the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach (https://www.gradeworkinggroup.org). The reviewer will consider five down-rating criteria: (1) risk of bias in individual study design, (2) inconsistency of results across studies, (3) indirectness of evidence (i.e., limitations in the generalisability of results), (4) imprecision (i.e., wide confidence intervals around estimates of effect), and (5) publication bias. Criteria for rating up may also be applied when there exist (1) large effect sizes, (2) a dose-response gradient, or (3) plausible residual confounding. After discussion between reviewers, the result will be categorised as very low, low, moderate, or high quality of evidence.

References:

An, X., Chai, C. S., Li, Y., Zhou, Y., & Yang, B. (2023). Modeling students' perceptions of artificial intelligence assisted language learning. *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2023.2246519

Chen, X., Xie, H., Zou, D., & Hwang, G.-J. (2020). Application and theory gaps during the rise of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, *1*, 1-20. https://doi.org/10.1016/j.caeai.2020.100002

Cherry, M. G., & Dickson, R. (2017). Defining My Review Question and Identifying Inclusion and Exclusion Criteria. In A. Boland, M. G. Cherry, & R. Dickson (Eds.), *Doing a Systematic Review: A Student's Guide* (2nd ed.) (pp. 68-86). SAGE.

Fleeman, N., & Dundar, Y. (2017). Data Extraction: Where Do I Begin? In A. Boland, M. G. Cherry, & R. Dickson (Eds.), *Doing a Systematic Review: A Student's Guide* (2nd ed.) (pp. 121-137). SAGE.

Fryer, L. K., Thompson, A., Nakao, K., Howarth, M., & Gallacher, A. (2020). Supporting self-efficacy beliefs and interest as educational inputs and outcomes: Framing AI and Human

partnered task experiences. *Learning and Individual Differences*, *80*, 1-11. https://doi.org/10.1016/j.lindif.2020.101850

Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., & Vedel, I. (2018). *Mixed methods appraisal tool (MMAT), version 2018*. Registration of Copyright (#1148552), Canadian Intellectual Property Office, Industry Canada. http://mixedmethodsappraisaltoolpublic.pbworks.com/w/file/fetch/127916259/MMAT_2018_crite ria-manual_2018-08-01_ENG.pdf

Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., & Vedel, I. (2020). *Reporting the results of the MMAT (version 2018)*. http://mixedmethodsappraisaltoolpublic.pbworks.com/w/file/fetch/140056890/Reporting%20the%20res ults%20of%20the%20MMAT.pdf

Ji, H., Han, I., & Ko, Y. (2023). A systematic review of conversational AI in language education: Focusing on the collaboration with human teachers. *Journal of Research on Technology in Education*, *55*(1), 48–63. https://doi.org/10.1080/15391523.2022.2142873

Liang, J.-C., Hwang, G.-J., Chen, M.-R. A., & Darmawansah, D. (2023). Roles and research foci of artificial intelligence in language education: An integrated bibliographic analysis and systematic review approach. *Interactive Learning Environments*, *31*(7), 4270–4296. https://doi.org/10.1080/10494820.2021.1958348

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence Unleashed: An argument for AI in education*. http://discovery.ucl.ac.uk/1475756/

Macaro, E. (2019). Systematic reviews in applied linguistics. In J. McKinley & H. Rose (eds.), *The Routledge Handbook of Research Methods in Applied Linguistics* (pp. 230–239). Routledge. https://doi.org/10.4324/9780367824471-20

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *BMJ*, *339*. https://doi.org/10.1136/bmj.b2535

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan -- a web and mobile app for systematic reviews. *Systematic Reviews*, 5(1), 1-10. https://doi.org/10.1186/s13643-016-0384-4

Pace, R., Pluye, P., Bartlett, G., Macaulay, A. C., Salsberg, J., Jagosh, J., & Seller, R. (2012). Testing the reliability and efficiency of the pilot Mixed Methods Appraisal Tool (MMAT) for systematic mixed studies review. *International Journal of Nursing Studies*, *49*(1), 47–53. https://doi.org/10.1016/j.ijnurstu.2011.07.002

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Blackwell Publishing. https://doi.org/10.1002/9780470754887

Schulz, J., Hamilton, C., Wonnacott, E., & Murphy, V. (2023). The impact of multi-word units in early foreign language learning and teaching contexts: A systematic review. *Review of Education, 11,* 1-26. https://doi. org/10.1002/rev3.3413

UNESCO. (2021). *AI and education: guidance for policy-makers*. https://unesdoc.unesco.org/ark:/48223/pf0000376709

UNESCO. (2023). *Guidance for generative AI in education and research*. https://unesdoc.unesco.org/ark:/48223/pf0000386693

Wang, J., Xie, H., Wang, F. L., Lee, L. K., & Au, O. T. S. (2021). Top-N personalized recommendation with graph neural networks in MOOCs. *Computers and Education: Artificial Intelligence, 2,* 1-10. https://doi.org/10.1016/j.caeai.2021.100010

Willis, S., Neil, R., Mellick, M. C., & Wasley, D. (2019). The Relationship Between Occupational Demands and Well-Being of Performing Artists: A Systematic Review. *Frontiers in Psychology, 10,* 1-20. https://doi.org/10.3389/fpsyg.2019.00393

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education, 16*(39), 1-27. https://doi.org/10.1186/s41239-019-0171-0

Zheng, L., Niu, J., Zhong, L., & Gyasi, J. F. (2021). The effectiveness of artificial intelligence on learning achievement and learning perception: A meta-analysis. *Interactive Learning Environments*, 1–15. https://doi.org/10.1080/10494820.2021.2015693

Zou, B., Liviero, S., Hao, M., & Wei, C. (2020). Artificial Intelligence Technology for EAP Speaking Skills: Student Perceptions of Opportunities and Challenges. In M. R. Freiermuth & N. Zarrinabadi (Eds.), *Technology and the Psychology of Second Language Learners and Users* (pp.433–463). Springer International Publishing. https://doi.org/10.1007/978-3-030-34212-8_17

Appendix 2. Sample Boolean search strings

Language	Database	Search string
English	British Education Index	TI (“artificial intelligence” OR AI OR “artificial neural network” OR “machine intelligence” OR “machine learning” OR “deep learning” OR robotic OR “intelligent support” OR “intelligent virtual reality” OR chatbot OR “chat bot” OR “conversational agent” OR "automated tutor" OR "personal tutor" OR "intelligent agent" OR "intelligent system" OR "intelligent tutor" OR "natural language processing" OR “generative artificial intelligence” OR “generative AI” OR GenAI OR “large language model” OR LLM OR ChatGPT OR GPT-3.5 OR GPT-4) AND TX ("language learn*" OR "language education" OR “second language” OR L2 OR “foreign language” OR pronunciation OR vocabulary OR grammar OR listening OR speaking OR reading OR writing) AND TX (outcome* OR achieve* OR perform* OR effect* OR impact*) Peer Reviewed; Publication Date: 19560101- AND Apply equivalent subjects
Chinese	China Academic Journals (CNKI) - Chinese	(Subject= 人工智能 + AI + 人工神经网络 + 机器智能 + 机器学习 + 深度学习 + 智能支持 + 智慧虚拟现实 + 聊天机器人 + 智能辅导 + 对话代理 + 智能代理 + 智能系统 + 智能导师 + 自然语言处理 + 生成式人工智能 + 生成式 AI + 大语言模型) AND (Subject= 语言学习 + 语言教育 + 二语 +外语 + 发音 + 读音 + 词汇 + 词语 + 语法 + 听力 + 口语 + 阅读 +写作) AND (Subject= 结果 + 成绩 + 表现 + 效果 + 影响)

Appendix 3. The data extraction form

Category	Item	Response*	Instructions
Article overview	Date form completed		Provide the date in the format of dd/mm/yyyy.
	Reference		List the full reference of the article.
	Document source		Clarify the source of the article (e.g., source database, website, etc.).
	Type of publication		Identify the type of the article (e.g., peer-reviewed journal article, thesis, book chapter, etc.)
	Language of publication		Indicate the language in which the article is written.
Study overview	Research questions		Present the research questions. Use quotation marks ("...") and include page numbers when they are directly quoted from the article.
	Study design		Identify if it is an RCT/ non-RCT/ cohort study/ cross-sectional survey study/ mixed methods study.
	Time of the study		Indicate when the study was conducted.
	Location of the study		Indicate where the study was conducted.
	Study setting		Identify if the study is conducted in laboratory, in-class, or out-of-class setting.
Participants	Source of recruitment		Summarise how the participants were recruited.
	Number of participants		(1) Record the total number if there is only one group of participants.
			(2) Record the number of participants in the experimental and the control group respectively.
	Age of participants		Present the mean or range of the participants' age.
	Gender of participants		Present descriptive statistics for the male and female participants.

	School setting		Indicate the educational level of the participants (e.g., primary school, middle school, university, etc.).
	Other socio-demographic information		Record any information about the participants (e.g., socioeconomic background).
	Languages spoken		List all languages spoken (L1/L2/…/Ln) and proficiency levels of each language.
	Target language		Write down the specific language (e.g., English, French, etc.).
	Target language skills		Write down the specific language skills (e.g., pronunciation, vocabulary, grammar, listening, speaking, reading, or writing).
Intervention	Specific intervention		What AI technology is used as the intervention?
	Number of the intervention group		Record how many intervention groups were involved.
	Duration of intervention		Record how long the intervention lasted in the format of xx month(s), xx week(s), xx day(s).
	Other relevant intervention details		Provide any detail that could be relevant to the review.
	Comparator		If there is a comparison group, record the comparative results of different groups.
Outcomes	Data type		Identify if quantitative, qualitative, or mixed data was collected.
	Outcome measurement		List the tests to measure learning outcomes.
	Specific outcomes		(1) Identify the specific dependent variable that is measured (e.g., test scores, learners' attitudes, etc.).
			(2) Record details, such as means, standard deviations, or qualitative data.
	Effect size		Record effect sizes, if any.

	Conclusion		Summarise the researcher's conclusions.
* Note down "N/A" if relevant information is not provided.			

Appendix 4. MMAT and explanations

Part I: Mixed Methods Appraisal Tool (MMAT), version 2018

Category of study designs	**Methodological quality criteria**	**Responses**			
		Yes	No	Can't tell	Comments
Screening questions (for all types)	S1. Are there clear research questions?				
	S2. Do the collected data allow to address the research questions?				
	Further appraisal may not be feasible or appropriate when the answer is 'No' or 'Can't tell' to one or both screening questions.				
1. Qualitative	1.1. Is the qualitative approach appropriate to answer the research question?				
	1.2. Are the qualitative data collection methods adequate to address the research question?				
	1.3. Are the findings adequately derived from the data?				
	1.4. Is the interpretation of results sufficiently substantiated by data?				
	1.5. Is there coherence between qualitative data sources, collection, analysis and interpretation?				
2. Quantitative randomized controlled trials	2.1. Is randomization appropriately performed?				
	2.2. Are the groups comparable at baseline?				
	2.3. Are there complete outcome data?				
	2.4. Are outcome assessors blinded to the intervention provided?				
	2.5 Did the participants adhere to the assigned intervention?				
3. Quantitative non-randomized	3.1. Are the participants representative of the target population?				
	3.2. Are measurements appropriate regarding both the outcome and intervention (or exposure)?				

	3.3. Are there complete outcome data?				
	3.4. Are the confounders accounted for in the design and analysis?				
	3.5. During the study period, is the intervention administered (or exposure occurred) as intended?				
4. Quantitative descriptive	4.1. Is the sampling strategy relevant to address the research question?				
	4.2. Is the sample representative of the target population?				
	4.3. Are the measurements appropriate?				
	4.4. Is the risk of nonresponse bias low?				
	4.5. Is the statistical analysis appropriate to answer the research question?				
5. Mixed methods	5.1. Is there an adequate rationale for using a mixed methods design to address the research question?				
	5.2. Are the different components of the study effectively integrated to answer the research question?				
	5.3. Are the outputs of the integration of qualitative and quantitative components adequately interpreted?				
	5.4. Are divergences and inconsistencies between quantitative and qualitative results adequately addressed?				
	5.5. Do the different components of the study adhere to the quality criteria of each tradition of the methods involved?				

Part II: MMAT Explanations

1. Qualitative studies	Methodological quality criteria
"Qualitative research is an approach for exploring and understanding the meaning individuals or groups ascribe to a social or human problem" (Creswell, 2013b, p. 3). Common qualitative research approaches include (this list if not exhaustive): **Ethnography** The aim of the study is to describe and interpret the shared cultural behaviour of a group of individuals. **Phenomenology** The study focuses on the subjective experiences and interpretations of a phenomenon encountered by individuals. **Narrative research** The study analyzes life experiences of an individual or a group. **Grounded theory** Generation of theory from data in the process of conducting research (data collection occurs first).	1.1. Is the qualitative approach appropriate to answer the research question? Explanations The qualitative approach used in a study (see non-exhaustive list on the left side of this table) should be appropriate for the research question and problem. For example, the use of a grounded theory approach should address the development of a theory and ethnography should study human cultures and societies. This criterion was considered important to add in the MMAT since there is only one category of criteria for qualitative studies (compared to three for quantitative studies).
	1.2. Are the qualitative data collection methods adequate to address the research question? Explanations This criterion is related to data collection method, including data sources (e.g., archives, documents), used to address the research question. To judge this criterion, consider whether the method of data collection (e.g., in depth interviews and/or group interviews, and/or observations) and the form of the data (e.g., tape recording, video material, diary, photo, and/or field notes) are adequate. Also, clear justifications are needed when data collection methods are modified during the study.
	1.3. Are the findings adequately derived from the data? Explanations This criterion is related to the data analysis used. Several data analysis methods have been developed and their use depends on the research question and qualitative approach. For example, open, axial and selective coding is often associated with grounded theory, and within- and cross-case analysis is often seen in case study.

Case study In-depth exploration and/or explanation of issues intrinsic to a particular case. A case can be anything from a decision-making process, to a person, an organization, or a country.	1.4. Is the interpretation of results sufficiently substantiated by data? Explanations The interpretation of results should be supported by the data collected. For example, the quotes provided to justify the themes should be adequate.
Qualitative description There is no specific methodology, but a qualitative data collection and analysis, e.g., in-depth interviews or focus groups, and hybrid thematic analysis (inductive and deductive). Key references: Creswell (2013a); Sandelowski (2010); Schwandt (2015)	1.5. Is there coherence between qualitative data sources, collection, analysis and interpretation? Explanations There should be clear links between data sources, collection, analysis and interpretation.

2. Quantitative randomized controlled trials	Methodological quality criteria
Randomized controlled clinical trial: A clinical study in which individual participants are allocated to intervention or control groups by randomization (intervention assigned by researchers). Key references: Higgins and Green (2008); Higgins et al. (2016); Oxford Centre for Evidence-based Medicine (2016); Porta et al. (2014)	2.1. Is randomization appropriately performed? Explanations In a randomized controlled trial, the allocation of a participant (or a data collection unit, e.g., a school) into the intervention or control group is based solely on chance. Researchers should describe how the randomization schedule was generated. A simple statement such as 'we randomly allocated' or 'using a randomized design' is insufficient to judge if randomization was appropriately performed. Also, assignment that is predictable such as using odd and even record numbers or dates is not appropriate. At minimum, a simple allocation (or unrestricted allocation) should be performed by following a predetermined plan/sequence. It is usually achieved by referring to a published list of random numbers, or to a list of random assignments generated by a computer. Also, restricted allocation can be performed such as blocked randomization (to ensure particular allocation ratios to the intervention groups), stratified randomization (randomization performed separately within strata), or minimization (to make small groups closely similar with respect to several characteristics). Another important characteristic to judge if randomization was appropriately performed is allocation concealment that protects assignment sequence until allocation. Researchers and participants should be unaware of the assignment sequence up to the point of allocation. Several strategies can be used to ensure allocation concealment such relying on a central randomization by a third party, or the use of sequentially numbered, opaque, sealed envelopes (Higgins et al., 2016).
	2.2. Are the groups comparable at baseline? Explanations Baseline imbalance between groups suggests that there are problems with the randomization. Indicators from baseline imbalance include: "(1) unusually large differences between intervention group sizes; (2) a substantial excess in statistically significant differences in baseline characteristics than would be expected by chance alone; (3) imbalance in key prognostic factors (or baseline measures of outcome variables) that are unlikely to be due to chance; (4) excessive similarity in baseline characteristics that is not compatible with chance; (5) surprising absence of one or more key characteristics that would be expected to be reported"

	(Higgins et al., 2016, p. 10).
	2.3. Are there complete outcome data? Explanations Almost all the participants contributed to almost all measures. There is no absolute and standard cut-off value for acceptable complete outcome data. Agree among your team what is considered complete outcome data in your field and apply this uniformly across all the included studies. For instance, in the literature, acceptable complete data value ranged from 80% (Thomas et al., 2004; Zaza et al., 2000) to 95% (Higgins et al., 2016). Similarly, different acceptable withdrawal/dropouts rates have been suggested: 5% (de Vet et al., 1997; MacLehose et al., 2000), 20% (Sindhu et al., 1997; Van Tulder et al., 2003) and 30% for a follow-up of more than one year (Viswanathan and Berkman, 2012).
	2.4. Are outcome assessors blinded to the intervention provided? Explanations Outcome assessors should be unaware of who is receiving which interventions. The assessors can be the participants if using participant reported outcome (e.g., pain), the intervention provider (e.g., clinical exam), or other persons not involved in the intervention (Higgins et al., 2016).
	2.5 Did the participants adhere to the assigned intervention? Explanations To judge this criterion, consider the proportion of participants who continued with their assigned intervention throughout follow-up. “Lack of adherence includes imperfect compliance, cessation of intervention, crossovers to the comparator intervention and switches to another active intervention.” (Higgins et al., 2016, p. 25).

3. Quantitative non-randomized studies	Methodological quality criteria
Non-randomized studies are defined as any quantitative studies estimating the effectiveness of an intervention or studying other exposures that do not use randomization to allocate units to comparison groups (Higgins and Green, 2008). Common designs include (this list if not exhaustive): **Non-randomized controlled trials** The intervention is assigned by researchers, but there is no randomization, e.g., a pseudo-randomization. A non- random method of allocation is not reliable in producing alone similar groups. **Cohort study** Subsets of a defined population are assessed as exposed, not exposed, or exposed at different degrees to factors of interest. Participants are followed over time to determine if an outcome occurs (prospective longitudinal). **Case-control study** Cases, e.g., patients, associated with a certain	3.1. Are the participants representative of the target population? Explanations Indicators of representativeness include: clear description of the target population and of the sample (inclusion and exclusion criteria), reasons why certain eligible individuals chose not to participate, and any attempts to achieve a sample of participants that represents the target population. 3.2. Are measurements appropriate regarding both the outcome and intervention (or exposure)? Explanations Indicators of appropriate measurements include: the variables are clearly defined and accurately measured; the measurements are justified and appropriate for answering the research question; the measurements reflect what they are supposed to measure; validated and reliability tested measures of the intervention/exposure and outcome of interest are used, or variables are measured using ‘gold standard’. 3.3. Are there complete outcome data? Explanations Almost all the participants contributed to almost all measures. There is no absolute and standard cut-off value for acceptable complete outcome data. Agree among your team what is considered complete outcome data in your field (and based on the targeted journal) and apply this uniformly across all the included studies. For example, in the literature, acceptable complete data value ranged from 80% (Thomas et al., 2004; Zaza et al., 2000) to 95% (Higgins et al., 2016). Similarly, different acceptable withdrawal/dropouts rates have been suggested: 5% (de Vet et al., 1997; MacLehose et al., 2000), 20% (Sindhu et al., 1997; Van Tulder et al., 2003) and 30% for follow-up of more than one year (Viswanathan and Berkman, 2012).

outcome are selected, alongside a corresponding group of controls. Data is collected on whether cases and controls were exposed to the factor under study (retrospective). **Cross-sectional analytic study** At one particular time, the relationship between health- related characteristics (outcome) and other factors (intervention/exposure) is examined. E.g., the frequency of outcomes is compared in different population subgroups according to the presence/absence (or level) of the intervention/exposure. Key references for non-randomized studies: Higgins and Green (2008); Porta et al. (2014); Sterne et al. (2016); Wells et al. (2000)	3.4. Are the confounders accounted for in the design and analysis? Explanations Confounders are factors that predict both the outcome of interest and the intervention received/exposure at baseline. They can distort the interpretation of findings and need to be considered in the design and analysis of a non-randomized study. Confounding bias is low if there is no confounding expected, or appropriate methods to control for confounders are used (such as stratification, regression, matching, standardization, and inverse probability weighting). 3.5 During the study period, is the intervention administered (or exposure occurred) as intended? Explanations For intervention studies, consider whether the participants were treated in a way that is consistent with the planned intervention. Since the intervention is assigned by researchers, consider whether there was a presence of contamination (e.g., the control group may be indirectly exposed to the intervention) or whether unplanned co-interventions were present in one group (Sterne et al., 2016). For observational studies, consider whether changes occurred in the exposure status among the participants. If yes, check if these changes are likely to influence the outcome of interest, were adjusted for, or whether unplanned co-exposures were present in one group (Morgan et al., 2017).

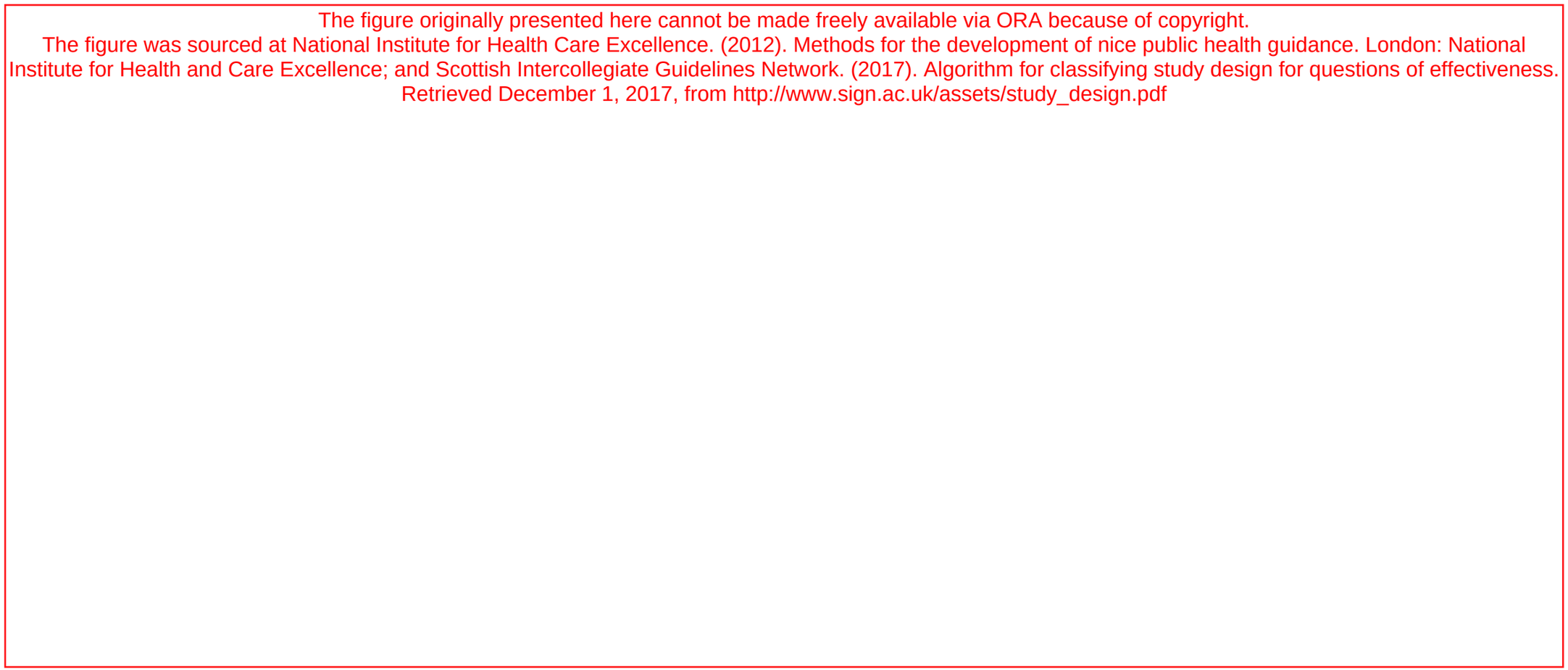
4. Quantitative descriptive studies	Methodological quality criteria
Quantitative descriptive studies are "concerned with and designed only to describe the existing distribution of variables without much regard to causal relationships or other hypotheses" (Porta et al., 2014, p. 72). They are used to monitoring the population, planning, and generating hypothesis (Grimes and Schulz, 2002). Common designs include the following single-group studies (this list if not exhaustive): **Incidence or prevalence study without comparison group** In a defined population at one particular time, what is happening in a population, e.g., frequencies of factors (importance of problems), is described (portrayed). **Survey** "Research method by which information is gathered by asking people questions on a specific topic and the data collection procedure is	4.1. Is the sampling strategy relevant to address the research question? Explanations Sampling strategy refers to the way the sample was selected. There are two main categories of sampling strategies: probability sampling (involve random selection) and non-probability sampling. Depending on the research question, probability sampling might be preferable. Non- probability sampling does not provide equal chance of being selected. To judge this criterion, consider whether the source of sample is relevant to the target population; a clear justification of the sample frame used is provided; or the sampling procedure is adequate. 4.2. Is the sample representative of the target population? Explanations There should be a match between respondents and the target population. Indicators of representativeness include: clear description of the target population and of the sample (such as respective sizes and inclusion and exclusion criteria), reasons why certain eligible individuals chose not to participate, and any attempts to achieve a sample of participants that represents the target population. 4.3. Are the measurements appropriate? Explanations Indicators of appropriate measurements include: the variables are clearly defined and accurately measured, the measurements are justified and appropriate for answering the research question; the measurements reflect what they are supposed to measure; validated and reliability tested measures of the outcome of interest are used, variables are measured using 'gold standard', or questionnaires are pre-tested prior to data collection.

standardized and well defined." (Bennett et al., 2011, p. 3). **Case series** A collection of individuals with similar characteristics are used to describe an outcome. **Case report** An individual or a group with a unique/unusual outcome is described in detail. Key references: Critical Appraisal Skills Programme (2017); Draugalis et al. (2008)	4.4. Is the risk of nonresponse bias low? Explanations Nonresponse bias consists of "an error of nonobservation reflecting an unsuccessful attempt to obtain the desired information from an eligible unit." (Federal Committee on Statistical Methodology, 2001, p. 6). To judge this criterion, consider whether the respondents and non- respondents are different on the variable of interest. This information might not always be reported in a paper. Some indicators of low nonresponse bias can be considered such as a low nonresponse rate, reasons for nonresponse (e.g., noncontacts vs. refusals), and statistical compensation for nonresponse (e.g., imputation). The nonresponse bias is might not be pertinent for case series and case report. This criterion could be adapted. For instance, complete data on the cases might be important to consider in these designs.
	4.5. Is the statistical analysis appropriate to answer the research question? Explanations The statistical analyses used should be clearly stated and justified in order to judge if they are appropriate for the design and research question, and if any problems with data analysis limited the interpretation of the results.

5. Mixed methods studies	Methodological quality criteria
Mixed methods (MM) research involves combining qualitative (QUAL) and quantitative (QUAN) methods. In this tool, to be considered MM, studies have to meet the following criteria (Creswell and Plano Clark, 2017): (a) at least one QUAL method and one QUAN method are combined; (b) each method is used rigorously in accordance to the generally accepted criteria in the area (or tradition) of research invoked; and (c) the combination of the methods is carried out at the minimum through a MM design (defined *a priori*, or emerging) and the integration of the QUAL and QUAN phases, results, and data. Common designs include (this list if not exhaustive): **Convergent design** The QUAL and QUAN components are usually (but not necessarily) concomitant. The purpose is to examine the same phenomenon by interpreting QUAL and QUAN results (bringing data analysis together at the interpretation stage), or by integrating QUAL and QUAN datasets (e.g., data on same cases), or by transforming data (e.g., quantization of qualitative data).	5.1. Is there an adequate rationale for using a mixed methods design to address the research question? Explanations The reasons for conducting a mixed methods study should be clearly explained. Several reasons can be invoked such as to enhance or build upon qualitative findings with quantitative results and vice versa; to provide a comprehensive and complete understanding of a phenomenon or to develop and test instruments (Bryman, 2006).
	5.2. Are the different components of the study effectively integrated to answer the research question? Explanations Integration is a core component of mixed methods research and is defined as the "explicit interrelating of the quantitative and qualitative component in a mixed methods study" (Plano Clark and Ivankova, 2015, p. 40). Look for information on how qualitative and quantitative phases, results, and data were integrated (Pluye et al., 2018). For instance, how data gathered by both research methods was brought together to form a complete picture (e.g., joint displays) and when integration occurred (e.g., during the data collection-analysis or/and during the interpretation of qualitative and quantitative results).
	5.3. Are the outputs of the integration of qualitative and quantitative components adequately interpreted? Explanations This criterion is related to meta-inference, which is defined as the overall interpretations derived from integrating qualitative and quantitative findings (Teddlie and Tashakkori, 2009). Meta-inference occurs during the interpretation of the findings from the integration of the qualitative and quantitative components, and shows the added value of conducting a mixed methods study rather than having two separate studies.

Sequential explanatory design Results of the phase 1 - QUAN component inform the phase 2 - QUAL component. The purpose is to explain QUAN results using QUAL findings. E.g., the QUAN results guide the selection of QUAL data sources and data collection, and the QUAL findings contribute to the interpretation of QUAN results. **Sequential exploratory design** Results of the phase 1 - QUAL component inform the phase 2 - QUAN component. The purpose is to explore, develop and test an instrument (or taxonomy), or a conceptual framework (or theoretical model). E.g., the QUAL findings inform the QUAN data collection, and the QUAN results allow a statistical generalization of the QUAL findings. Key references: Creswell et al. (2011); Creswell and Plano Clark, (2017); O'Cathain (2010)	5.4. Are divergences and inconsistencies between quantitative and qualitative results adequately addressed? Explanations When integrating the findings from the qualitative and quantitative components, divergences and inconsistencies (also called conflicts, contradictions, discordances, discrepancies, and dissonances) can be found. It is not sufficient to only report the divergences; they need to be explained. Different strategies to address the divergences have been suggested such as reconciliation, initiation, bracketing and exclusion (Pluye et al., 2009b). Rate this criterion 'Yes' if there is no divergence. 5.5. Do the different components of the study adhere to the quality criteria of each tradition of the methods involved? Explanations The quality of the qualitative and quantitative components should be individually appraised to ensure that no important threats to trustworthiness are present. To appraise 5.5, use criteria for the qualitative component (1.1 to 1.5), and the appropriate criteria for the quantitative component (2.1 to 2.5, or 3.1 to 3.5, or 4.1 to 4.5). The quality of both components should be high for the mixed methods study to be considered of good quality. The premise is that the overall quality of a mixed methods study cannot exceed the quality of its weakest component. For example, if the quantitative component is rated high quality and the qualitative component is rated low quality, the overall rating for this criterion will be of low quality.

Part III: Algorithm for selecting the study categories to rate in the MMAT*



*Adapted from National Institute for Health Care Excellence. (2012). *Methods for the development of nice public health guidance*. London: National Institute for Health and Care Excellence; and Scottish Intercollegiate Guidelines Network. (2017). *Algorithm for classifying study design for questions of effectiveness*. Retrieved December 1, 2017, from http://www.sign.ac.uk/assets/study_design.pdf.

Appendix 5. Detailed information for each study

Note 1: NS = The information was "not specified" in the paper.
Note 2: M = Male; F = Female

ID	Reference	Educational level	Sample size	Gender	Country	Type of AI technology	L1	L2/FL	L2 proficiency	Language skills	Study design	Publication type	Theoretical background	RoB
1	Chia & Li (2020)	Secondary	88	NS	China	ITS	Chinese	Japanese	Beginning	Grammar	Quantitative non-RCT study	Journal article	No theory	Moderate
2	Chew & Chua (2020)	Tertiary	6	NS	NS	ITS	NS	Chinese	Mixed	Vocabulary	Qualitative case study	Journal article	Engagement Theory	Moderate
3	Ayedoun et al. (2019)	Tertiary	40	24 M 16 F	Japan	Chatbot	Japanese	English	NS	Speaking	Mixed methods study	Journal article	No theory	Low
4	Zhou (2023)	Tertiary	356	234 M 116 F	China	AI-assisted language learning platform	Chinese	English	Mixed	Integrated	Quantitative cohort study	Journal article	No theory	Moderate
5	Li et al. (2023)	Secondary	4	2 M 2 F	United States	Chatbot	English	Chinese	Mixed	Writing	Mixed methods study	Journal article	No theory	Moderate
6	Qasem et al. (2023)	Tertiary	40	40 M	Saudi Arabia	Chatbot	Arabic	English	NS	Vocabulary	Mixed methods study	Journal article	The Noticing Hypothesis	Moderate
7	Kim (2022b)	Tertiary	119	50 M 69 F	Korea	AI-assisted language learning platform	Korean	English	Mixed	Integrated	Quantitative non-RCT study	Journal article	No theory	Moderate
8	El Shazly (2021)	Tertiary	48	10 M 38 F	Egypt	Chatbot	Arabic	English	Intermediate	Speaking	Mixed methods study	Journal article	Interaction Hypothesis	Moderate
9	Duong & Suppasetseree (2024)	Tertiary	30	13 M 17 F	Vietnam	Chatbot	Vietnamese	English	Mixed	Speaking	Mixed methods study	Journal article	Connectivism Theory	Moderate
10	Agnes & Srinivasan (2024)	Tertiary	60	NS	India	AI-assisted language learning application	NS	English	Intermediate	Vocabulary	Quantitative non-RCT study	Journal article	Cognitive Load Theory	Low
11	Zhang & Huang (2024)	Secondary	52	25 M 27 F	China	Chatbot	Chinese	English	NS	Vocabulary	Mixed methods study	Journal article	No theory	Moderate

12	Dhanapal et al. (2024)	Mixed	72	28 M 44 F	NS	AI-assisted language learning application	NS	English	Mixed	Integrated	Mixed methods study	Journal article	No theory	Moderate
13	Meyer et al. (2024)	Secondary	459	184 M 240 F 35 Nonbinary or unknown	Germany	AI-assisted writing and evaluation tool	NS	English	NS	Writing	Quantitative RCT study	Journal article	Self-Determination Theory	Moderate
14	Wiboolyasarin et al. (2024)	Tertiary	39	7 M 32 F	Thailand	Chatbot	NS	Thai	Mixed	Writing	Quantitative non-RCT study	Journal article	Sociocultural Theory	Moderate
15	Wale (2024)	Tertiary	92	NS	Ethiopia	AI-assisted writing and evaluation tool	NS	English	Upper-intermediate	Writing	Mixed methods study	Journal article	No theory	Moderate
16	Jamshed et al. (2024)	Tertiary	258	171 M 87 F	Saudi Arabia	AI-assisted language learning application	Arabic	English	NS	Integrated	Quantitative cross-sectional study	Journal article	No theory	Moderate
17	Kemelbekova et al. (2024)	Tertiary	51	NS	Kazakhstan	Chatbot	NS	English	Intermediate	Speaking	Mixed methods study	Journal article	No theory	Moderate
18	Lee et al. (2024)	Tertiary	80	34 M 46 F	Korea	AI-assisted writing and evaluation tool	Korean	English	Mixed	Writing	Mixed methods study	Journal article	No theory	Moderate
19	Darejeh et al. (2024)	Secondary	14	14 M	Iran	ITS	NS	English	NS	Integrated	Quantitative RCT study	Conference proceeding	No theory	Low
20	Oraif (2024)	Tertiary	6	6 F	Saudi Arabia	Speech recognition tool	NS	English	Advanced	Integrated	Qualitative case study	Journal article	No theory	Low
21	Guo & Wen (2023)	Elementary	37	NS	Singapore	AI-assisted language learning platform	English	Chinese	Mixed	Vocabulary	Mixed methods study	Conference proceeding	No theory	Moderate
22	Madhavi et al. (2023)	Secondary	100	NS	India	AI-assisted language learning application	NS	English	NS	Speaking	Quantitative non-RCT study	Conference proceeding	No theory	Moderate
23	Nozhovnik et al. (2023)	Tertiary	55	NS	Ukraine	Chatbot	NS	English	Mixed	Integrated	Mixed methods study	Journal article	No theory	Low
24	Zou et al. (2023)	Tertiary	70	NS	China	AI-assisted language learning application	NS	English	NS	Speaking	Mixed methods study	Journal article	No theory	Moderate

25	Al Mahmud (2023)	Secondary	156	77 M 79 F	Saudi Arabia	AI-assisted writing and evaluation tool	Arabic	English	Mixed	Writing	Mixed methods study	Journal article	No theory	Moderate
26	Khampusaen et al. (2023)	Mixed	15	4 M 11 F	Thailand	AI-assisted language learning application	Isan	English	Mixed	Pronunciation	Mixed methods study	Journal article	Sociocultural Theory	Low
27	Markus et al. (2023)	Tertiary	96	NS	Russia	AI-assisted language learning application	NS	English	Mixed	Integrated	Mixed methods study	Conference proceeding	Technology Acceptance Model	Moderate
28	El-Zeiny et al. (2023)	Tertiary	49	NS	Egypt	Speech recognition tool	Malaysian	Arabic	Intermediate	Pronunciation	Quantitative RCT study	Conference proceeding	No theory	Moderate
29	Liao (2023)	Tertiary	93	NS	China	ITS	Chinese	English	Mixed	Listening	Mixed methods study	Conference proceeding	No theory	High
30	Kim & Cha (2023)	Tertiary	113	39 M 74 F	Korea	AI-assisted language learning application	Korean	English	Mixed	Reading	Mixed methods study	Journal article	No theory	Low
31	Vu et al. (2022)	Tertiary	53	NS	Vietnam	AI-assisted language learning application	Vietnamese	English	Mixed	Listening	Quantitative RCT study	Journal article	Input Hypothesis	High
32	Kim (2022a)	Tertiary	486	NS	Korea	AI-assisted language learning application	Korean	English	Mixed	Integrated	Quantitative RCT study	Journal article	No theory	Low
33	Wang (2022)	Tertiary	178	62 M 116 F	China	AI-assisted writing and evaluation tool	Chinese	English	Mixed	Writing	Mixed methods study	Journal article	Expectancy Disconfirmation Theory	Moderate
34	Liu & Ardakani (2022)	Tertiary	30	24 M 6 F	China	ITS	Chinese	English	Mixed	Integrated	Quantitative non-RCT study	Journal article	No theory	Moderate
35	Gómez-Ortiz & Buentello-Montoya (2022)	Mixed	40	20 M 20 F	Mexico	AI-assisted language learning application	Spanish	Polish	Mixed	Integrated	Mixed methods study	Conference proceeding	No theory	Moderate
36	Lee & Jeon (2022)	Elementary	67	30 M 37 F	Korea	Chatbot	Korean	English	NS	Integrated	Qualitative description study	Journal article	No theory	Low
37	Kim et al. (2022)	Tertiary	121	NS	Korea	ITS	Korean	English	NS	Integrated	Quantitative RCT study	Journal article	No theory	Moderate
38	Wu (2022)	Tertiary	60	NS	China	AI-assisted language learning application	Chinese	English	Mixed	Speaking	Mixed methods study	Conference proceeding	No theory	Moderate

39	Kharis (2022)	Tertiary	36	NS	Indonesia	Chatbot	Indonesian	German	Beginning	Grammar	Mixed methods study	Journal article	No theory	Moderate
40	Gayed et al. (2022)	Tertiary	10	NS	Japan	AI-assisted writing and evaluation tool	Japanese	English	Mixed	Writing	Quantitative non-RCT study	Journal article	No theory	Moderate
41	Sun (2022)	Tertiary	154	NS	China	AI-assisted writing and evaluation tool	Chinese	English	Mixed	Writing	Mixed methods study	Conference proceeding	No theory	Moderate
42	Çakmak (2022)	Tertiary	90	33 M 57 F	Turkey	Chatbot	Turkish	English	Intermediate	Speaking	Mixed methods study	Journal article	No theory	Moderate
43	Luo & Liang (2022)	Tertiary	157	NS	China	AI-assisted writing and evaluation tool	Chinese	English	Mixed	Writing	Mixed methods study	Conference proceeding	No theory	Moderate
44	Srinivasan & Murthy (2021)	Elementary		NS	India	ITS	NS	English	Mixed	Reading	Quantitative RCT study	Journal article	No theory	Low
45	Nong et al. (2021)	Elementary	86	NS	China	AI-assisted language learning application	Chinese	English	NS	Integrated	Quantitative RCT study	Conference proceeding	No theory	Low
46	Li & Peng (2021)	Tertiary	59	15 M 44 F	China	AI-assisted language learning platform	Chinese	English	Intermediate	Listening	Mixed methods study	Conference proceeding	No theory	Moderate
47	Li & Graesser (2021)	Mixed	93	62 M 31 F	India	ITS	NS	English	NS	Writing	Quantitative RCT study	Journal article	No theory	Moderate
48	Kim et al. (2021a)	Tertiary	49	49 F	Korea	Chatbot	Korean	English	Mixed	Speaking	Mixed methods study	Journal article	No theory	Moderate
49	Nghi et al. (2019)	Tertiary	200	NS	Vietnam	Chatbot	Vietnamese	English	Mixed	Grammar	Quantitative RCT study	Journal article	No theory	Low
50	Tegos et al. (2014)	Tertiary	30	4 M 26 F	Ukraine	ITS	Mixed	English	NS	Vocabulary	Mixed methods study	Journal article	Sociocultural Theory	Moderate
51	Nagata (1995)	Tertiary	18	NS	United States	ITS	NS	Japanese	NS	Grammar	Quantitative RCT study	Journal article	No theory	Low
52	Shaikh et al. (2023)	Tertiary	10	8 M 2 F	Norway	Chatbot	NS	English	Mixed	Integrated	Quantitative survey study	Journal article	No theory	Low
53	Kim & Su (2024)	Tertiary	65	NS	China	Chatbot	Chinese	Korean	NS	Speaking	Mixed methods study	Journal article	WTC theory	Moderate
54	Qiao & Zhao (2023)	Tertiary	93	41 M 52 F	China	AI-assisted language learning application	Mixed	English	Mixed	Speaking	Quantitative RCT study	Journal article	Social Constructivist Theory	Low
55	Karataş et al. (2024)	Tertiary	13	3 M 10 F	Turkey	Chatbot	Turkish	English	Mixed	Integrated	Qualitative case study	Journal article	Engagement Theory	Moderate

56	Xiao & Zhi (2023)	Tertiary	5	NS	China	Chatbot	Chinese	English	NS	Integrated	Qualitative case study	Journal article	No theory	Moderate
57	Fu et al. (2020)	Tertiary	260	127 M 133 F	China	Speech recognition tool	Chinese	English	NS	Speaking	Mixed methods study	Journal article	Affordance Theory	Moderate
58	Shafiee Rad (2024)	Tertiary	66	26 M 40 F	Iran	AI-assisted language learning application	Persian	English	Intermediate	Speaking	Mixed methods study	Journal article	Sociocultural Theory	Low
59	Kwon et al. (2023)	Elementary	75	38 M 37 F	Korea	Chatbot	Mixed	English	NS	Writing	Quantitative RCT study	Journal article	No theory	Moderate
60	Song & Song (2023)	Tertiary	50	NS	China	Chatbot	Chinese	English	Mixed	Writing	Mixed methods study	Journal article	Social Constructivist Theory	Low
61	Mohamed & Alian (2023)	Secondary	64	21 M 43 F	Egypt	Chatbot	Arabic	English	NS	Integrated	Quantitative survey study	Journal article	No theory	Moderate
62	Chen et al. (2022)	Tertiary	2	NS	China	ITS	Chinese	English	Intermediate	Integrated	Qualitative case study	Journal article	Engagement Theory	Moderate
63	Fryer et al. (2019)	Tertiary	91	69 M 22 F	Japan	Chatbot	NS	English	Mixed	Integrated	Mixed methods study	Journal article	No theory	Moderate
64	Lee et al. (2023)	Elementary	121	64 M 57 F	Korea	AI-assisted language learning application	Korean	English	Mixed	Reading	Mixed methods study	Journal article	No theory	Moderate
65	Chen et al. (2024)	Secondary	70	36 M 34 F	China	AI-assisted writing and evaluation tool	Chinese	English	Mixed	Writing	Quantitative RCT study	Journal article	No theory	Low
66	Ye et al. (2022)	Secondary	56	NS	China	Chatbot	Chinese	English	Mixed	Speaking	Quantitative RCT study	Journal article	No theory	Moderate
67	Yuan (2023)	Elementary	74	NS	China	Chatbot	Chinese	English	Intermediate	Speaking	Mixed methods study	Journal article	WTC theory	Moderate
68	Yang et al. (2022)	Mixed	314	214 M 95 F 5 Unknown	Korea	Chatbot	Korean	English	Mixed	Speaking	Mixed methods study	Journal article	No theory	Moderate
69	Peña-Acuña & Crismán-Pérez (2022)	Tertiary	128	37 M 91 F	Spain	AI-assisted language learning application	Spanish	English	Intermediate	Integrated	Mixed methods study	Journal article	No theory	Moderate
70	Wei (2023)	Tertiary	60	24 M 36 F	China	AI-assisted language learning application	Chinese	English	Intermediate	Integrated	Mixed methods study	Journal article	Social Constructivist Theory	Low
71	Zhang et al. (2024)	Tertiary	131	32 M 99 F	China	AI-assisted language learning application	Chinese	English	Mixed	Speaking	Quantitative non-RCT study	Journal article	Control-Value Theory	Moderate

72	Fathi et al. (2024)	Tertiary	65	32 M 33 F	Iran	Chatbot	Mixed	English	Mixed	Speaking	Mixed methods study	Journal article	Engeström's activity theory	Moderate
73	Chen et al. (2020)	Tertiary	58	27 M 31 F	China	Chatbot	Mixed	Chinese	Mixed	Vocabulary	Mixed methods study	Journal article	Technology Acceptance Model	Moderate
74	Lin & Mubarok (2021)	Tertiary	50	NS	China	Chatbot	NS	English	Mixed	Speaking	Mixed methods study	Journal article	No theory	Moderate
75	Jeon (2023)	Elementary	53	30 M 23 F	Korea	Chatbot	Korean	English	Beginning	Vocabulary	Mixed methods study	Journal article	Sociocultural Theory	Moderate
76	Boudouaia et al. (2024)	Tertiary	76	NS	Algeria	Chatbot	Arabic	English	Intermediate	Writing	Quantitative RCT study	Journal article	Technology Acceptance Model	Low
77	Zhang et al. (2023b)	Tertiary	30	5 M 25 F	China	Chatbot	Chinese	English	Upper-intermediate	Writing	Mixed methods study	Journal article	No theory	Moderate
78	Zhang et al. (2023c)	Tertiary	15	3 M 12 F	China	Chatbot	Chinese	English	Upper-intermediate	Writing	Mixed methods study	Journal article	No theory	Moderate
79	Liu et al. (2023)	Tertiary	103	28 M 75 F	China	AI-assisted writing and evaluation tool	NS	English	Mixed	Writing	Mixed methods study	Journal article	No theory	Moderate
80	Shafiee Rad & Roohani (2024)	Mixed	48	48 F	Iran	AI-assisted language learning application	Persian	English	Intermediate	Pronunciation	Mixed methods study	Journal article	Social Constructivist Theory	Moderate
81	Feng & Wang (2023)	Elementary	85	44 M 41 F	China	ITS	Chinese	English	Mixed	Reading	Mixed methods study	Journal article	Input Hypothesis	Moderate
82	T.-C. Hsu et al. (2023)	Elementary	47	26 M 21 F	China	AI-assisted language learning application	Chinese	English	NS	Vocabulary	Quantitative non-RCT study	Journal article	No theory	Moderate
83	Fryer et al. (2017)	Tertiary	122	NS	Japan	Chatbot	NS	English	NS	Speaking	Quantitative RCT study	Journal article	No theory	Moderate
84	Hwang et al. (2023)	Tertiary	104	NS	Indonesia	AI-assisted writing and evaluation tool	NS	English	Mixed	Writing	Mixed methods study	Journal article	No theory	Moderate
85	Ghafouri et al. (2024)	Tertiary	48	19 M 29 F	Iran	Chatbot	Persian	English	Intermediate	Writing	Quantitative RCT study	Journal article	Social-Cognitive Theory	Moderate
86	Shafiee Rad et al. (2023)	Tertiary	46	NS	Iran	AI-assisted writing and evaluation tool	Persian	English	Upper-intermediate	Writing	Mixed methods study	Journal article	Social Constructivist Theory	Moderate

87	Liu & Chen (2023)	Elementary	72	NS	China	AI-assisted language learning application	Chinese	English	Mixed	Vocabulary	Quantitative non-RCT study	Journal article	Dual-Coding Theory	Moderate
88	Bao (2019)	Mixed	40	10 M 27 F 3 Unknown	Thailand	Chatbot	Mixed	English	Mixed	Speaking	Quantitative non-RCT study	Journal article	No theory	High
89	Kostikova & Miasoiedova (2019)	Tertiary	32	NS	Ukraine	AI-assisted writing and evaluation tool	NS	English	Upper-intermediate	Writing	Quantitative non-RCT study	Journal article	No theory	Moderate
90	Azamatova et al. (2023)	Tertiary	64	NS	Kazakhstan	AI-assisted language learning application	NS	Russian	NS	Integrated	Quantitative RCT study	Journal article	No theory	Low
91	Huang & Wang (2023)	Tertiary	132	NS	China	ITS	Chinese	French	Mixed	Vocabulary	Quantitative RCT study	Journal article	No theory	Low
92	Chang et al. (2021)	Tertiary	53	9 M 44 F	China	AI-assisted writing and evaluation tool	Chinese	English	Intermediate	Writing	Mixed methods study	Conference proceeding	Technology Acceptance Model	Moderate
93	M.-H. Hsu et al. (2023)	Tertiary	48	NS	China	Chatbot	Chinese	English	Beginning	Speaking	Quantitative RCT study	Journal article	No theory	Moderate
94	Ghafouri (2024)	Secondary	30	19 M 11 F	Iran	Chatbot	NS	English	Intermediate	Integrated	Quantitative RCT study	Journal article	No theory	Low
95	Wang & Feng (2023)	Tertiary	83	34 M 49 F	China	Chatbot	Chinese	English	NS	Reading	Mixed methods study	Conference proceeding	No theory	Moderate
96	Kim et al. (2021b)	Tertiary	110	NS	Korea	Chatbot	Korean	English	Mixed	Speaking	Mixed methods study	Journal article	No theory	Moderate
97	Muniandy & Selvanathan (2024)	Tertiary	40	16 M 24 F	Malaysia	Chatbot	NS	English	Mixed	Speaking	Mixed methods study	Journal article	Technology Acceptance Model	Moderate
98	Mahapatra (2024)	Tertiary	134	NS	India	Chatbot	NS	English	NS	Writing	Mixed methods study	Journal article	No theory	Low
99	Liu et al. (2024)	Elementary	95	NS	China	Chatbot	Chinese	English	NS	Reading	Quantitative survey study	Journal article	No theory	Moderate
100	Chien et al. (2022)	Secondary	73	36 M 37 F	China	Chatbot	Chinese	English	NS	Integrated	Quantitative non-RCT study	Journal article	No theory	Moderate
101	Wale & Kassahun (2024)	Tertiary	92	NS	Ethiopia	AI-assisted writing and evaluation tool	NS	English	Intermediate	Writing	Mixed methods study	Journal article	Sociocultural theory	Moderate

102	Liu et al. (2024)	Elementary	30	NS	China	Chatbot	Chinese	English	NS	Reading	Mixed methods study	Journal article	No theory	Moderate
103	Zhang (2024)	Tertiary	66	NS	China	AI-assisted writing and evaluation tool	Chinese	English	Mixed	Writing	Mixed methods study	Journal article	Technological Pedagogical Content Knowledge	Moderate
104	Liu et al. (2022)	Elementary	68	NS	China	Chatbot	Chinese	English	Beginning	Reading	Mixed methods study	Journal article	No theory	Moderate
105	Davis (2022)	Mixed	36	NS	United States	AI-assisted language learning application	Chinese	English	Mixed	Integrated	Mixed methods study	Doctoral dissertation	Interaction Hypothesis	Moderate

Appendix 6. Risk of bias of individual studies

ID	Reference			1. Qualitative					2. Quantitative RCTs					3. Quantitative non-randomized					4. Quantitative descriptive					5. Mixed methods					Quality
		S1	S2	1.1	1.2	1.3	1.4	1.5	2.1	2.2	2.3	2.4	2.5	3.1	3.2	3.3	3.4	3.5	4.1	4.2	4.3	4.4	4.5	5.1	5.2	5.3	5.4	5.5	
1	Chia & Li (2020)																												***
2	Chew & Chua (2020)																												***
3	Ayedoun et al. (2019)																												****
4	Zhou (2023)																												***
5	Li et al. (2023)																												***
6	Qasem et al. (2023)																												***
7	Kim (2022b)																												***
8	El Shazly (2021)																												***
9	Duong & Suppasetseree (2024)																												***
10	Agnes & Srinivasan (2024)																												****
11	Zhang & Huang (2024)																												***
12	Dhanapal et al. (2024)																												***
13	Meyer et al. (2024)																												***
14	Wiboolyasarin et al. (2024)																												***
15	Wale (2024)																												***
16	Jamshed et al. (2024)																												***
17	Kemelbekova et al. (2024)																												***
18	Lee et al. (2024)																												***

19	Darejeh et al. (2024)																												****
20	Oraif (2024)																												*****
21	Guo & Wen (2023)																												***
22	Madhavi et al. (2023)																												***
23	Nozhovnik et al. (2023)																												****
24	Zou et al. (2023)																												***
25	Al Mahmud (2023)																												***
26	Khampusaen et al. (2023)																												****
27	Markus et al. (2023)																												***
28	El-Zeiny et al. (2023)																												***
29	Liao (2023)																												*
30	Kim & Cha (2023)																												****
31	Vu et al. (2022)																												**
32	Kim (2022a)																												****
33	Wang (2022)																												***
34	Liu & Ardakani (2022)																												***
35	Gómez-Ortiz & Buentello-Montoya (2022)																												***
36	Lee & Jeon (2022)																												*****
37	Kim et al. (2022)																												***
38	Wu (2022)																												***
39	Kharis (2022)																												***
40	Gayed et al. (2022)																												***

41	Sun (2022)																													***
42	Çakmak (2022)																													***
43	Luo & Liang (2022)																													***
44	Srinivasan & Murthy (2021)																													*****
45	Nong et al. (2021)																													****
46	Li & Peng (2021)																													***
47	Li & Graesser (2021)																													***
48	Kim et al. (2021a)																													***
49	Nghi et al. (2019)																													****
50	Tegos et al. (2014)																													***
51	Nagata (1995)																													****
52	Shaikh et al. (2023)																													*****
53	Kim & Su (2024)																													***
54	Qiao & Zhao (2023)																													****
55	Karataş et al. (2024)																													***
56	Xiao & Zhi (2023)																													***
57	Fu et al. (2020)																													***
58	Shafiee Rad (2024)																													****
59	Kwon et al. (2023)																													***
60	Song & Song (2023)																													****
61	Mohamed & Alian (2023)																													***
62	Y.-L. Chen et al. (2022)																													***

63	Fryer et al. (2019)																												***
64	Lee et al. (2023)																												***
65	Chen et al. (2024)																												****
66	Ye et al. (2022)																												***
67	Yuan (2023)																												***
68	Yang et al. (2022)																												***
69	Peña-Acuña & Crismán-Pérez (2022)																												***
70	Wei (2023)																												****
71	Zhang et al. (2024)																												***
72	Fathi et al. (2024)																												***
73	H. Chen et al. (2020)																												***
74	Lin & Mubarok (2021)																												***
75	Jeon (2023)																												***
76	Boudouaia et al. (2024)																												****
77	Zhang et al. (2023b)																												***
78	Zhang et al. (2023c)																												***
79	Liu et al. (2023)																												***
80	Shafiee Rad & Roohani (2024)																												***
81	Feng & Wang (2023)																												***
82	T.-C. Hsu et al. (2023a)																												***
83	Fryer et al. (2017)																												***
84	Hwang et al. (2023)																												***

85	Ghafouri et al. (2024)																												***
86	Shafiee Rad et al. (2023)																												***
87	Liu & Chen (2023)																												***
88	Bao (2019)																												**
89	Kostikova & Miasoiedova (2019)																												***
90	Azamatova et al. (2023)																												****
91	Huang & Wang (2023)																												****
92	Chang et al. (2021)																												***
93	M.-H. Hsu et al. (2023b)																												***
94	Ghafouri (2024)																												****
95	Wang & Feng (2023)																												***
96	Kim et al. (2021b)																												***
97	Muniandy & Selvanathan (2024)																												***
98	Mahapatra (2024)																												****
99	Liu et al. (2024)																												***
100	Chien et al. (2022)																												***
101	Wale & Kassahun (2024)																												***
102	Liu et al. (2024)																												***
103	Zhang (2024)																												***
104	Liu et al. (2022)																												***
105	Davis (2022)																												***