ISSN 2201-2982
2025/2

IELTS Research Reports **Online Series**

Exploring the possibilities of integrating communicative AI into the IELTS test preparation process

Carlo Perrotta, Ute Knoch, Neil Selwyn and Sima Mohammadi

IELTS Research

BRITISH COUNCIL idp CAMBRIDGE English

IELTS is jointly owned by the British Council; IDP IELTS; and Cambridge University Press & Assessment

Exploring the possibilities of integrating communicative AI into the IELTS test preparation process

This report examines Generative AI (GenAI) from a perspective of language education and language test preparation. While there is already a growing body of research on GenAI and language learning, the role of prompting, understood as a novel form of human-computer communication, has been neglected. To address this gap, we carried out a study in two parts.

Funding

This research was funded by the IELTS Partners: British Council, IDP IELTS, and Cambridge University Press & Assessment. Grant awarded 2023.

Publishing details

Published by the IELTS Partners: British Council, IDP IELTS, and Cambridge University Press & Assessment © 2025.

How to cite this report

Perrotta, C., Knoch, U., Selwyn, N. and Mohammadi, S. (2025). Exploring the possibilities of integrating communicative AI into the IELTS test preparation process *IELTS Research Reports Online Series*, No. 2/25. British Council, IDP IELTS, and Cambridge University Press & Assessment.
Available at: https://ielts.org/researchers/our-research/research-reports

Introduction

This study by Perrotta, Knoch, Selwyn and Mohammadi was conducted with support from the IELTS Partners (British Council, IDP: IELTS Australia, and Cambridge University Press & Assessment), as part of the IELTS joint-funded research program. Research funded by the British Council and IDP: IELTS Australia under this program complement those conducted or commissioned by Cambridge University Press & Assessment, and together inform the ongoing validation and improvement of IELTS.

A significant body of research has been produced since the joint-funded research program started in 1995, with over 200 empirical studies receiving grant funding. After undergoing a process of peer reviews and revision, many of the studies have been published in academic journals, in several IELTS-focused volumes in the *Studies in Language Testing* series (http://www.cambridgeenglish.org/silt), and in the *IELTS Research Reports* series. Since 2012, to facilitate timely access, the research reports have been published on the IELTS website immediately after completing the peer review and revision process.

The rapid development of Generative AI (GenAI) has sparked significant interest across various domains, including second language education and assessment. While much of the existing research thus far has explored how GenAI can facilitate language learning, one crucial aspect has remained under explored: prompting. As a new mode of human-computer communication, prompting plays a vital role in shaping the interactions and outputs of AI systems, yet its pedagogical and practical implications have not been extensively explored thus far. This research sought to address that gap by investigating the role of prompting in the context of English language learning and language test preparation (specifically IELTS test preparation).

This mixed-methods study was conducted in two parts. First, a scoping review of relevant, emerging literature was carried out to explore how prompting GenAI language models can support language learning and assessment. This review identified three key application scenarios—text generation, test item creation, and automated assessment—analysing the construction and replicability of prompts within each. Additionally, this review sheds light on the emerging significance of 'meta prompts' which operate invisibly in the AI's back-end, contrasting with user-generated prompts.

The second strand of this study focused on qualitative, ethnographic fieldwork in a language school in a major Australian city. The goal of this was to examine how AI prompting practices are developing in real educational settings according to real educational needs. Through interviews and observations, this research strand illustrates how teachers and students engage in situated forms of sensemaking and prompting, often driven by practical needs and informal theories about AI. Notably, the researchers observed a pattern of 'tactical' prompting among students—a form of simple, yet purpose-driven, prompting which reflected immediate learning goals.

This study highlights the potential of GenAI prompting in supporting language learning and test preparation, including applications relevant to the IELTS test. However, as the researchers here recognise, issues such as inconsistency, bias, and limited contextual awareness will persist. While effective prompting can mitigate these problems, it requires a shift from viewing prompts as simple technical instructions to understanding them as communicative acts shaped by context, culture, and user intention.

The research presented here argues for a nuanced, literacy-based approach that supports meaningful, context-sensitive interactions. In terms of both learning and assessment, there is a case to be made for personalised, communicative approaches to AI that reflect the diverse realities of language education as it becomes a feature of everyday life.

NICK GLASSON
SENIOR RESEARCH MANAGER
CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

Exploring the possibilities of integrating communicative AI into the IELTS test preparation process

Abstract

The new horizon of human-computer communication and the changing nature of language education.

This report examines Generative AI (GenAI) from a perspective of language education and language test preparation. While there is already a growing body of research on GenAI and language learning, the role of prompting, understood as a novel form of human-computer communication, has been neglected. To address this gap, we carried out a study in two parts.

In the first part, we conducted a scoping review that focused on how prompting GenAI language models can support English language learning and assessment. The review identified three application scenarios for prompts: text generation, test item generation and automated assessment. For each scenario, we examined how prompts were constructed and how they could be replicated. In addition to these scenarios, the review also highlighted the emerging importance of 'meta prompts' which are distinct from user-oriented prompts in that they operate in the back-end of AI models and are not visible or modifiable. We also found that effective prompts can certainly increase the sophistication and precision of human-AI interactions, but the outputs of these interactions still display limitations in terms of contextual awareness, bias, reliability and performance consistency.

In the second part, we conducted qualitative fieldwork to understand how prompting as a communicative practice is emerging in a real context of language learning and language test preparation. To this end, we carried out interviews and observations in a language school located in a large Australian city. We found that both teachers and students engage in highly contextual forms of sensemaking that influence informal theories about GenAI use and prompting. We also found evidence of a form of prompting amongst students that we termed 'tactical': unsophisticated but reflecting pragmatic and subjective priorities.

In our conclusion, we reflect on the significance of prompting as pragmatic communication and suggest some implications and future research directions.

Authors' biodata

Dr Carlo Perrotta is Associate Professor of Digital Education at the Faculty of Education, University of Melbourne. His research examines the role of technology in multiple aspects of education, from pedagogy to policy. Carlo's recent work focuses on automation and artificial intelligence (AI) and his articles have been published in several field-leading journals at the intersection of education studies, technology-enhanced learning and media studies. His research has been funded by large international bodies such as the Australian Research Council, the European Commission, and the UK's ESRC.

Professor Ute Knoch directs the Language Testing Research Centre at the University of Melbourne. Her research focuses on language assessment for academic and professional purposes, assessment policy, rater-mediated assessment, and placement testing. Her published works include *Diagnostic Writing Assessment* (2009), *Fairness, Justice and Language Assessment* (2019), *Assessing English for Professional Purposes* (2020), *Scoring Second Language Performances* (2021), and *The Handbook of Language Assessment Across Modalities* (2022). Ute served as ALTAANZ co-president (2015–2016) and is current vice-president. She has also held ILTA Executive Board positions (2011–2014, 2017–2019).

Professor Neil Selwyn is a full-time professor in the Monash Faculty of Education and a global leader in education and technology research. His work has extensive reach among international policy, professional, and industry audiences, making him one of Australia's most-cited education researchers. The *Australian* newspaper's Research 2023 report identified his work as leading one of the 'top ten challenges' for Australian researchers, recognising him as 'a leading international researcher in digital education used in schools, universities and adult learning'. Over 25 years, Neil has pioneered education and technology studies, conducting early empirical research on digital innovations from internet-based learning to datafication.

Sima Mohammadi is the research assistant employed to work on this project. She is a PhD student in Education at Deakin University and holds a Master's degree in Applied Linguistics. Her research interests involve technology-enhanced learning, computer-supported language learning, and self-regulated learning. Sima has published in peer-reviewed journals, including *Learning Environments Research* and *TESL-EJ*.

Table of contents

List of tables

List of figures

1 Introduction

Large language models (LLMs) are a form of Generative AI (GenAI) trained on large amounts of textual data extracted from the Internet. They can produce a limitless range of textual outputs as well as engage in natural language interactions in an adaptive and emergent fashion. Language models underpin conversational agents or 'chatbots' such as ChatGPT, which has gained global prominence since its release in 2022, as well as similar tools such as Google's Gemini, Microsoft's CoPilot, Anthropic's Claude and, as of February 2025, DeepSeek.

The rapid and disruptive ascendance of GenAI has led to discussions about its possible role in education and its potential in language learning (Xu & Li, 2023). As models improve and begin to appear to excel in forms of sophisticated reasoning and creativity that used to be the preserve of humans (Hubert et al., 2024; Rodrigues et al., 2024), the scientific consensus is converging around two overarching principles which have multiple educational ramifications.

1. A transition from production expertise to evaluation expertise

While models may outperform humans in tasks that can be learned through pattern recognition across large datasets – like language translation, mathematical computation, or identifying visual patterns – they still struggle with tasks that require contextual and relational expertise. For example, while an AI language model can engage in conversations that can assist different aspects of language learning, it lacks the embodied experience of managing a language classroom, where relationships with students are built over time, and where teaching strategies reflect subtle social dynamics. Such contextual and relational expertise is now more relevant than ever, and it is called upon to evaluate the procedural dynamics and the outputs of GenAI in terms of appropriateness. In other words, as GenAI becomes increasingly commonplace, we are beginning to observe scenarios in which people are less involved in the production of knowledge as they spend time evaluating GenAI's outputs and adapting them to their contexts (Bearman et al., 2024).

2. The importance of natural language prompts

It is becoming apparent that the 'quality' of GenAI's outputs is dependent to a considerable degree on the quality of the natural language prompts inserted as inputs. These inputs play three key roles: a) priming and conditioning the model at the outset, b) steering and moderating the outputs in real time, and c) providing positive and negative reinforcement which contributes to the fine-tuning of the model.

These two principles also inform the current debate about the potential of GenAI for language education. Early research suggests the positive impacts of GenAI tools on different language skills, such as writing, communication aptitude, and vocabulary acquisition. Moreover, GenAI may also support language teaching and assessment by enabling the automated generation of scoring tools, lesson plans, and teaching materials (Law, 2024). Alongside this early optimism, there are also calls for caution.

For example, some concerns are beginning to emerge around the uneven impacts of AI on learning: performance in certain tasks (e.g., essay writing) may improve but evaluative knowledge does not, and indeed learners' dependence on this new technology may trigger 'metacognitive laziness', which can be defined as 'learners' dependence on AI assistance, offloading metacognitive load and less effectively associating responsible metacognitive processes with learning tasks' (Fan et al., 2024, p. 506). These processes are critical for self-regulation and include orientation, planning, monitoring, and evaluation. Compounding these issues are concerns that assessment is becoming increasingly unable to separate human performance from AI's performance (Barrot, 2023).

Against this background, the present study seeks to understand the extent to which purposeful prompting of a LLM can assist English language learning and assessment. Throughout this paper we will refer to 'prompting' as the main concept being explored. Our use of the term prompting incorporates two interrelated definitions: prompt engineering and prompt literacy. The former has become established in the mainstream discourse as a novel form of technical expertise, which involves strategies used in the input statements to get better results from LLMs (OpenAI, n.d). The process of crafting such input statements through natural language is described as 'iterative and interactive — a dialogue between humans and AI in an act of co-creation' (Oppenlaender et al., 2024, pp. 1-2). The latter notion of 'prompt literacy' features mainly in a smaller educational literature where the technicist implications of 'engineering' are downplayed and more emphasis is placed on the novel communicative and textual skills which can be mastered by non-experts, that is, 'individuals without formal instruction concerning AI and LLMs' (Knoth et al., 2024, p. 2).

Using this broad definition of prompting, the study explored two research questions:

RQ1: How is GenAI prompting being used in research to assist English language learning and assessment?

RQ2: How is prompting for GenAI being used in naturalistic conditions?

These two questions complement each other as they enable a comprehensive examination that includes the analysis of research trends and a focus on the authentic conditions of GenAI adoption. Considering the scholarly debates on the impact of GenAI tools in education, the findings of this research can provide a systematic and empirical grounding for the current discourses around the role of GenAI tools in English language learning. These insights can have implications for language learning, teaching, and language assessment practices and policies.

Each question is answered in a separate part of this report. To answer RQ1, we conducted a scoping review of the emerging literature that examines the role of prompting in English language learning and assessment. The time range we considered is 2021–2024. To answer RQ2, we carried out an ethnographic case study in a language school. The fieldwork lasted 11 months (from February 2024 to December 2024) and consisted of interviews with teachers and students, and observations of instances of prompting during workshops and classroom observations.

2. RQ1: How is GenAI prompting being used in research to assist English language learning and assessment?

2.1 Related work: what are the main approaches to prompting beyond language learning?

Before examining studies on prompting in the context of language learning, we immersed ourselves in the broader literature on prompting as a distinct area of research. This is a fast-evolving area of investigation which emerged very rapidly in the wake of GenAI and its purpose is to investigate the potentials and limits of natural language for machine instruction. We found that the following seven methods of prompting recur throughout the literature to date and can be considered as foundational in providing a basis for more context-specific approaches. The list does not claim to be exhaustive and is best viewed as a partial synthesis developed by the research team, albeit supported by some of the most highly cited studies in this novel area of research and professional practice.

1. Input-output (IO) prompting: this the most common way to query a language model, involving simple and direct requests without too much preparation. These simple requests can be further divided in zero-shot prompts, where no examples are provided for the LLM, and few-shot prompts, which include one or more examples or 'shots' for the LLM to learn from and follow to produce an output (Brown et al., 2020; Sahoo et al., 2024; Yang et al., 2023)

2. Chain-of-thought (CoT) prompting: instructing the LLM to think step-by-step to explain the analytical and reasoning process through which the output is produced. Asking the model to 'role play' a particular persona has been found to be an effective trigger for the CoT process (Kong et al., 2023; Wang et al., 2023; Wei et al., 2022).

3. Self-consistency with CoT, which seeks to improve upon CoT by instructing the model to consider different reasoning approaches for the same problem, and then select the most frequent (X. Wang et al., 2022).

4. Tree of thought prompting, in which a model does not simply rely on heuristics to choose an output (e.g., the most frequent solution), but is instructed to decompose a problem or task and explore all possible steps associated with each possible solution along different logical branches (Yao et al., 2023).

5. Progressive hint prompting, which involves crafting a relatively simple IO prompt and then providing incremental hints that guide the model towards a more fine-tuned output. The method shares some similarities with chain-of-thought but it is more interactive. While CoT typically asks for all reasoning steps at the outset, progressive hint prompting affords a degree of course correction and clarification after each step (Zhang et al., 2023)

6. Retrieval-Augmented Generation (RAG) is a further advancement which enhances a model's capabilities by instructing it to retrieve specific and accurate knowledge to support the generation of an output. For this method to be effective, the model must be integrated with a particular database or repository, or have direct access to the web (Lewis et al., 2020).

7. Meta-prompting, which is when prompting begins to operate at a higher level of abstraction. Three prompting approaches in this line can be observed across the literature: self-instruct (Y. Wang et al., 2022), chain-of-thought coding language (Chan et al., 2025), and metacognitive prompting for LLMs (Wang & Zhao, 2024). In a self-instruct prompt framework, prompts are not used as direct instructions to generate content, but as overarching guidelines for how a model is expected to behave. This is sometimes called 'bootstrapping' a model (Y. Wang et al., 2022). Chain-of-Thought Coding Language (CoT-CL), is when coding language is integrated into the prompt to improve the quality of output. In this way, the LLM is prompted to emulate the verification stages in the coding language using the CoT approach within a separate window and apart from the internal generation process of the LLM. Lastly, metacognitive prompting is to prompt the LLM to emulate and automate the process of comprehension clarification, judgement, critical evaluation, decision confirmation, and confidence assessment.

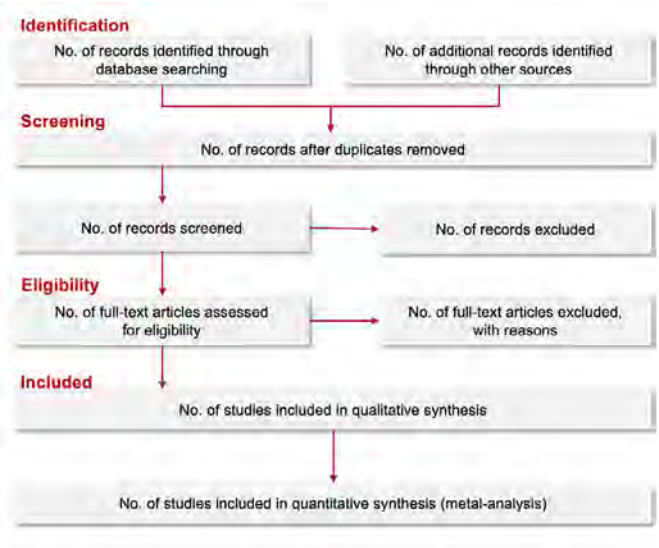
It must be stressed that the above list does not represent a hierarchy, as the effectiveness of a method will always depend on several factors such as the nature of the task and the evaluation criteria. For example, a simple Input-Output (IO) prompt might be suited to a straightforward task, while a Chain-of-Thought (CoT) one will add unnecessary reasoning layers. Similarly, a Tree of Thought prompt might be suited to problem-solving tasks but will be less effective – possibly counterproductive – when applied to creative tasks. Elements from different methods can also be combined to develop a great number of variations, which reflect multiple preferences and inclinations. Indeed, prompting has rapidly become a highly marketable form of expertise with scores of strategies, approaches and techniques vying for visibility and relevance. Furthermore, not all the above methods are user-oriented, with more sophisticated methods like RAG and meta-prompting explicitly targeting software developers and other professional users, who work on the back-end technical processes of LLMs that are inaccessible to front-end users. For instance, meta-prompting is an essential step in the creation of custom agents which are then released or marketed in specific settings, and it can be used as a form of guardrail operating in the back-end of prompts to ensure that a system complies with content and ethical criteria (Bai et al., 2022).

While these different prompting approaches suggest a growing knowledge of how to interact productively with GenAI – both in academia and non-specialist settings – there is a need to better understand how prompting can be used as a communicative strategy to interact with AI for more specific educational purposes, such as English language learning. These considerations were instrumental in defining the methodological features of our scoping review, which will be described in the next section.

2.2 Scoping review methodology

This scoping review is based on the PRISMA framework (Liberati et al., 2009), which provides a transparent and replicable protocol. This protocol is usually visually represented as a flow diagram (see Figure 1).

Figure 1: *The PRISMA four-phase flow diagram (Moher et al., 2009)*



The protocol is applicable to different types of literature reviews and meta-analyses so it must be adapted to specific research questions and contextual constraints and opportunities. The PRISMA protocol has been used in a great number of scoping as well as systematic literature reviews across multiple disciplines, including education research (Crompton & Burke, 2023; Santomauro et al., 2021; Shibuya et al., 2022).

2.3 Source selection

The PRISMA guidelines recommend using more than one database to avoid selection biases during the development of the literature corpus. For this reason, we used the same search query in Scopus and Web of Science (WoS). Scopus and WoS are currently the two leading citation databases, widely used across knowledge domains for academic research (Zhu & Liu, 2020).

2.4 Search criteria

The identification and inclusion of articles were guided by the following search criteria, which were used to construct a search query.

- LLM term: the article's title, abstract, author keywords or full text to have at least one of the following terms: large language model, generative language model, generative AI.
- Language learning and assessment term: article's title, abstract, author keywords or full text to have at least one of the following terms: language education, English as foreign language, English as second language, English as additional language, language testing, language assessment, Speaking, Listening, Writing, Reading.
- Prompting term: article's title, abstract, author keywords or full text to have at least one of the following terms: prompt engineering, prompting strategy, and prompting.

After an initial exploratory phase when we tested the query across databases, a search was carried out on 7 August 2024 in Scopus and WoS. The results were refined in terms of a) document type and b) limiting the subject area to social sciences, arts and humanities, psychology, and multidisciplinary. In addition, it was decided to exclude computational linguistics articles from the results as these studies proved to be largely technical and focused on specific aspects of linguistics rather than empirical investigations of prompts in a use-case scenario.

It was also decided to exclude image generation as the focus of this report is on large language models rather than other types of AI models. This search resulted in 1,618 items in Scopus and 147 items in WoS. This discrepancy reflects the fact that these two databases have different journal coverage, with WoS being particularly selective (Singh et al., 2021). Overall, 1,765 items were retrieved from databases.

2.5 Corpus screening

The process of evaluating the relevance of each article was conducted manually by two independent raters. When the eligibility of a particular item was uncertain, consensus was reached by re-reading that article in depth.

The inclusion criteria were as follows:

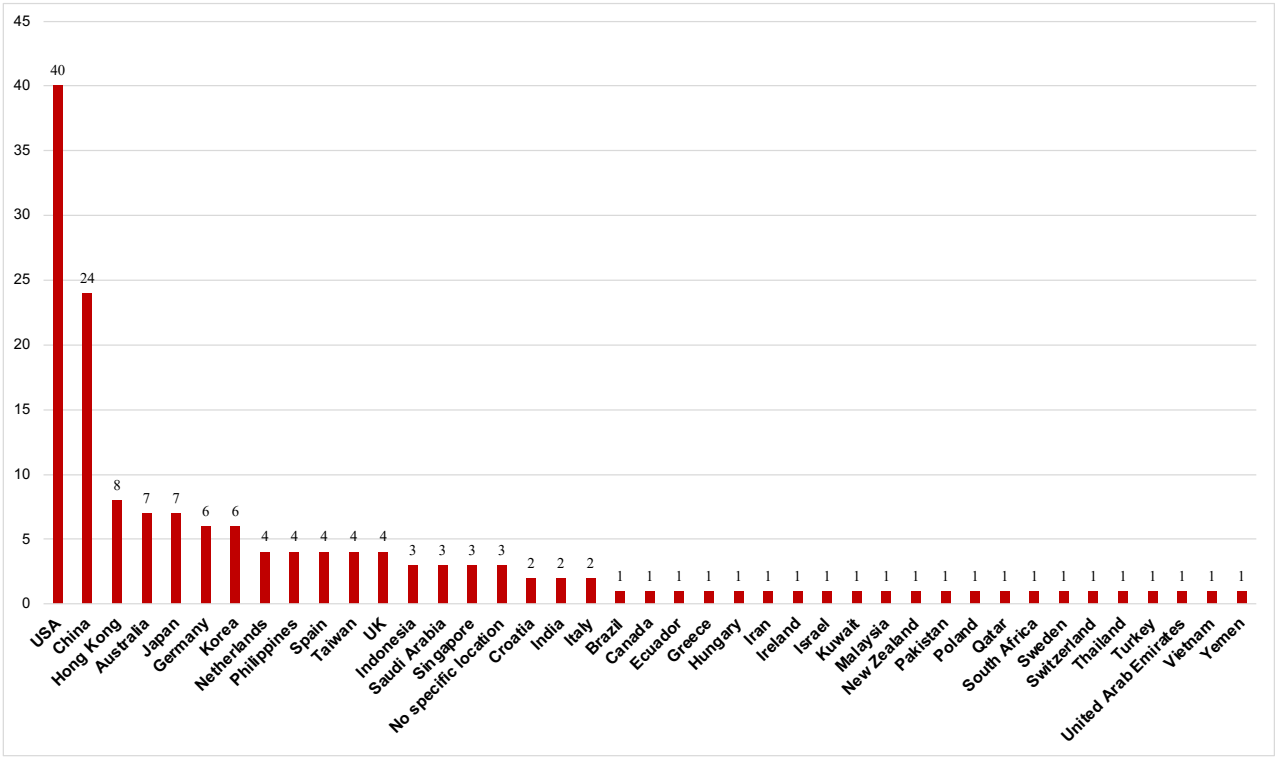
- peer reviewed: the item is from a peer-reviewed journal or peer-reviewed conference proceedings
- topic-relevance: the item is relevant to the application of LLMs in the context of language education; prompting has to be a substantive component of the study
- language: English
- date range: 2021–2024.

The exclusion criteria were as follows:

- articles written in a language other than English
- no-match topic: item is irrelevant to the application of LLMs in the context of language education.

The procedure led to a long list of 158 papers relevant to prompting for language education. The resulting list was examined to gauge the geographic distribution of prompting research. This revealed articles on prompting to be based primarily in the USA (N=40) and China (N=24) (Figure 2).

Figure 2: *Geographic distribution of publications*



Three articles did not specify a geographic location. Interestingly, these studies were conducted by corporate research and development (R&D) groups: META AI Research, Duolingo and Scispace. Prompted by this finding, we looked for additional industry affiliations across the corpus as we thought this was an interesting trend to document. Overall, there were seven studies carried out by authors affiliated with technology companies or testing organisations:

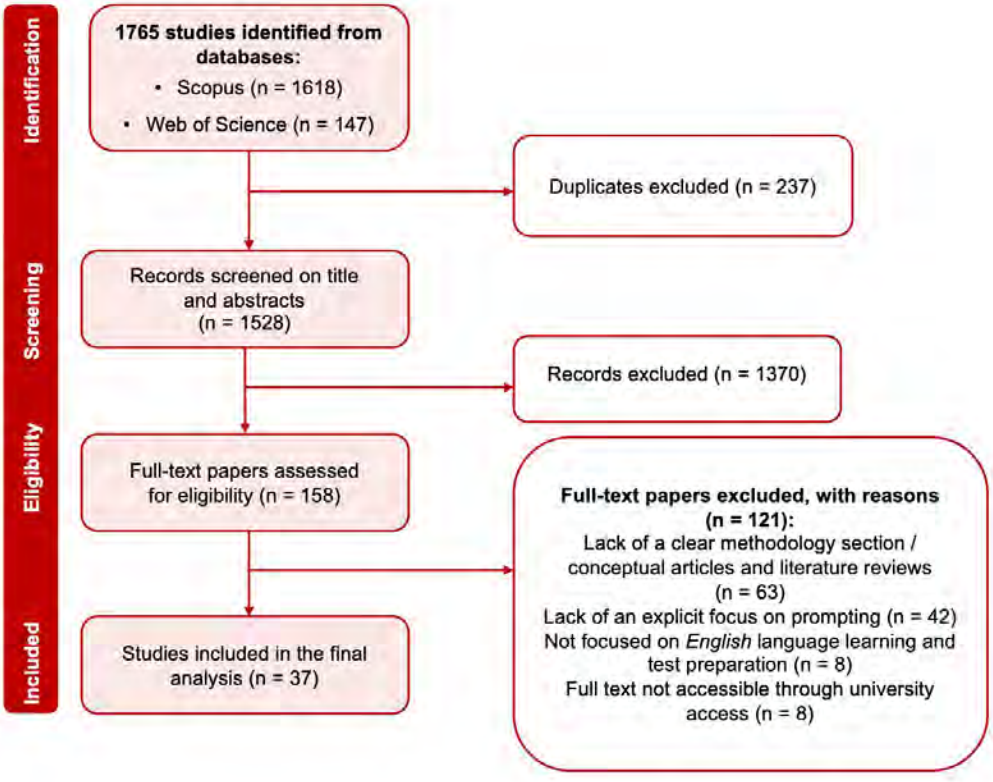
a) Meta AI research (Schick et al., 2022)
b) Duolingo (Yancey et al.)
c) Scispace (Roy et al., 2024)
d) ZUP Innovation (Pinto et al., 2023)
e) Cambium Assessment (Lottridge et al., 2024)
f) British Council (O'Sullivan, 2023)
g) Microsoft Azure AI (Zhang et al., 2023).

Following this preliminary mapping exercise which serves merely as background, the long list was refined further using the following stricter exclusion criteria, which allowed us to select only studies that had a distinct and deliberate empirical focus on prompting:

- lack of a clear methodology section focused on prompting (n =105)
- not focused on English language learning (N = 8)
- full text not accessible through university access (N = 8).

This fine-grained screening led to a final list of 37 items. The entire screening process that led to the final shortlist is summarised in Figure 3.

Figure 3: *PRISMA flow diagram of the scoping review*



2.6 Results of in-depth analysis of the shortlisted articles

In this section, we provide a critical analysis of the articles included in the final list (N=37) to outline the scope of research on prompting in English language learning and assessment in response to the research question of the study: *How is GenAI prompting being used in research to assist English language learning and assessment?*

2.6.1 What are the main trends and research interests across the studies?

All 37 studies included in the shortlist are focused on examining the challenges and opportunities of AI-assisted or fully automated language learning and assessment systems. The overarching trend is an interest in evaluating the performance of LLMs, either against specific quality criteria or in comparison with human performance. This trend translated into varied approaches to data generation. Of the 37 articles, 17 studies used human participants in the generation of empirical data – comparing students' and teachers' outputs with that of LLMs. For instance, Zhou et al. (2023) compared ChatGPT and Chinese intermediate English language learners' narrative writing. In another study, Bucol and Sangkawong (2024) compared ChatGPT's automated assessment with teachers' ratings.

In 20 studies, no human participants were involved directly in the generation of empirical data. In some cases, human experts were used ex-post to evaluate the quality of the AI output (e.g., Lee et al., 2023; Z. Wang et al., 2022). In one study, the LLM's automated assessment and feedback was compared with a pre-existing human-produced corpora of graded English essays (Mizumoto & Eguchi, 2023).

In examining the research designs, a distinction was also noted between studies with prompts written directly using a Chatbot (N=20) – ChatGPT being the most common – and studies where researchers engaged in a more sophisticated fashion with an AI model's Application Programming Interface (N=17), for optimising assessment, feedback, test questions, and prompt templates (Table 1).

Table 1: *How studies engaged with LLMs (Chatbots vs. API)*

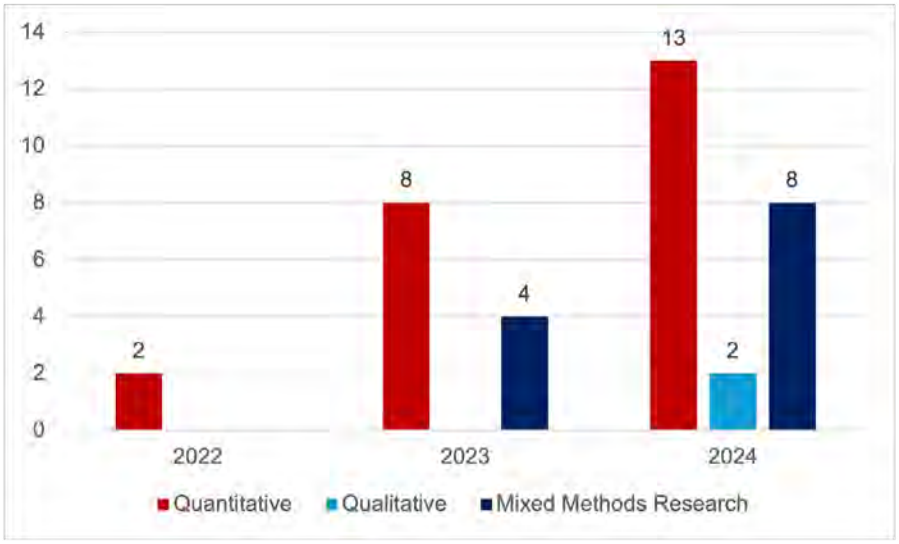
API integration and use	**17**
Claude 2	1
GPT 3	5
GPT 3.5	5
GPT 4	6
InstructGPT	2
Indication of API use (ChatGPT)	2
Chatbot	**20**
Bard	1
Bing chat	2
ChatGPT	21
Llama	1
PEER	1
POE	1

Engagement with a model through an API is considered salient as it affords a higher degree of control over the model rather than relying on the generic settings that regulate a Chatbot's behaviour. API-level tweaking is also a necessary step when 'integrating' a LLM into an existing platform or app to create customised conversational agents (e.g., Liu et al., 2023; Loem et al., 2023; Mizumoto et al., 2024). Working with APIs enables access to a range of settings that can shape the generation of prompts and outputs. These settings include, for example, the 'temperature' of a model, which makes its responses more or less deterministic, or more or less stochastic. In other words, operating at the API level has a direct impact on model behaviour and therefore should be reported as a key factor in the methodology. Our review showed that while a few studies have reported such API-level decisions, (Wang & Gayed, 2024; Yancey et al.), other studies have either not explained these settings or have partially reported them (e.g., Mizumoto & Eguchi, 2023). This highlights the importance of transparency when reporting research on prompting, because application and use of APIs impact comparability of results.

2.6.2 What are the main methods used?

The analysis of methodological designs in the final list showed that studies have mainly used quantitative (N = 23) and mixed-method research (N = 12) designs, respectively. Overall, only two studies in 2024 used qualitative designs (Koltovskaia et al., 2024; Liu et al., 2024). Figure 4 shows the trend across the 2022–2024 period of the adopted methodologies, suggesting the dominance of quantitative approaches in the context of experimental, comparative, and LLM evaluation studies where a combination of statistical techniques, machine learning and natural language processing approaches were applied to existing corpora of essays for automated scoring or other purposes.

Figure 4: *Methodological approaches in the reviewed studies*



As already mentioned, only two studies used fully qualitative designs. These studies examined students' behavioural and cognitive engagement (Koltovskaia et al., 2024) and cognitive processes (Liu et al., 2024) in the use of prompts for language learning. Other studies used mixed-methods (e.g., Lin & Chen, 2024; Woo et al., 2024) and quantitative designs (e.g., Meyer et al., 2024; Wang & Gayed, 2024).

2.6.3 How are prompts used to support language learning and assessment?

Overall, the reviewed studies tended to use prompts in the context of writing, with less focus on reading, listening, and speaking skills. Specifically, 32 studies investigated writing skills (e.g.,Dang et al., 2023; Escalante et al., 2023; Guo & Wang, 2023; Han et al., 2023; Liu et al., 2024; Meyer et al., 2024; L. Wang et al., 2024; Zhou et al., 2023), one explored listening skills (Aryadoust et al., 2024), four examined reading comprehension (Bezirhan & Von Davier, 2023; Lee et al., 2023; Lin & Chen, 2024; X. Wang et al., 2024), and one study explored pragmatic competence in text generation of conversations and dialogues (Chen et al., 2024).

We identified three application scenarios for the use of prompts. Sometimes different applications were present in the same study: text generation (N=10); test item generation (N=4); automated assessment (N=20). In addition to these scenarios, we identified cases in which prompts did not have an immediate applied dimension in terms of user-AI interaction but were instead used as unseen instructions operating in the background, enabling the creation of custom or fine-tuned AI agents (N=6). For each of these categories, we will now describe the strategies underpinning prompt design and provide examples of how some prompts can be developed. This interpretative synthesis provides a systemic and empirically grounded account of how prompts can be used in the context of English language learning and assessment. However, we should caution the reader that the true extent to which these prompt templates are valid and reliable for the scenarios described previously requires further exploration and validation in empirical settings.

2.6.3.1 Text generation

Prompts in this category can be used to replicate and model writing genres, such as narrative writing (Bai et al., 2024; Zhou et al., 2023), argumentative essays and multimodal PowerPoint using ChatGPT and Bing Chat (Liu et al., 2024) ,and letter writing (Jovic & Mnasri, 2024; Woo et al., 2024). Prompts can be productively designed for text generation through the careful integration in the prompt of instructions relating to elements of the desired output. For instance, the prompt might instruct the model to integrate specific narrative conventions and roles such as characters, setting, plot, and lessons learned. Alternatively, the prompt might require compliance with certain linguistic criteria or testing specifications. This method could be used by teachers and students to model writing and testing competencies, experimenting with how different combinations of elements work together and produce different outcomes – two examples are provided in Table 2.

Table 2: *Prompting for text generation*

Overall strategy	Prompt elements	Prompt example
Input-output	Story elements: Characters, setting, plot, and lessons/moral Digital story elements: Dramatic question, point of view, and pacing Story outline: Opening conversation, multiple-choice questions, and ending conversation	"Write me an outline of a story about [character(s) and rough setting and plot]. You need to include [lesson(s)]. The story should consist of five parts with a beginning, climax, and ending." (adapted from Bai et al. 2024)
Input-output	Linguistic components: syntactic simplicity, word concreteness, referential cohesion, and deep cohesion	"Write a personal experience in 500 words, then revise it in terms of syntactic simplicity and deep cohesion." (adapted from Zhou et al. 2023)

Input-output with zero-shot and few-shot variations	Fine-tuning components: Number of 'shots' (examples): zero-shot, one-shot and few-shots.	Zero-shot variation "This is an informative story generator. Generate an informative story about bees [for a 10-year-old]. It includes sections about bees' bodies, their honey production, social life and importance to the ecosystem. The sections should be informative and engaging [for a 10-year-old]." One-shot variation "This is an informative story generator. Generate an informative story about bees [for a 10-year-old]. It includes sections about bees' bodies, their honey production, social life and importance to the ecosystem. The sections should be informative and engaging [for a 10-year-old]. See example below based on ants. Story: Ants Small and strong Lift up a rock, and a family of ants might be crawling there. Ants are small insects, but they are very strong. Ants have six strong legs that help them carry big loads such as sticks and other insects. They can lift 20 times their own body weight." (adapted from Bezirhan and Von Davier, 2023)

2.6.3.2 Test item generation

Prompts can be used to develop a range of test materials such as generic reading comprehension questions (Lin & Chen, 2024; X. Wang et al., 2024), and discipline-specific comprehension questions (Z. Wang et al., 2022). An interesting example is offered by Aryadoust et al (2024), who used this approach to generate IELTS test items (listening scripts coupled with multiple-choice questions) for test-takers across a continuum of proficiency levels (academic, low, intermediate, and advanced). Another notable example is the generation of reading comprehension questions based on task types (i.e., literal and inferential) and the format of questions (wh-questions, cloze test, and yes/no questions) with two options of open-ended or multiple-choice questions (Lee et al., 2023).

The development of prompts for test item generation follows the same logic described previously for text generation. The inclusion of specific pre-defined, research-based elements in the prompt is essential, but given the precise nature of test items, more attention must be paid to the content-specificity of examples and the skills/outcomes of a well-defined target population (Table 3).

Table 3: *Prompting for test item generation*

Overall strategy	Prompt elements	Prompt example
Progressive-hint prompting	Listening script elements: the length of the script, the target language use domain (e.g., academic proficiency listening test), the topic, and proficiency of target audience, authentic real world linguistic features ("disfluency" features like "uh" and "uhm"). Multiple-choice questions (MCQ) elements: parameters for each MCQ defining measurement criteria for various aspects of the listening script, such as the main idea, purpose, information details, and inferred information.	Listening script prompts Initial prompt: "Write a 300–400 words text for an academic proficiency listening test. The text should be a university lecture on the topic of X." Subsequent hints (examples): "Rewrite the following listening script for adult English language learners with a low/intermediate/advanced proficiency level in English." "Incorporate natural disfluency features ("uh" and "uhm")." MCQ prompts Initial prompt: "write an MCQ item to assess comprehension of the main idea of the text." Subsequent hint prompts (examples) "Options must NOT use the words as the text so use paraphrases." 'The MCQ must not be answerable unless the test-taker listens to the text.' (Adapted from Aryadoust et al. 2024)
Input-output, role-play	Specific skills that need to be assessed: comprehension, summarisation, inferencing, etc.	"Take the role of an English language test developer. Based on the given reading passage, please generate 9 multiple-choice items. One for determining the meaning of words (phrases) from the context, two for understanding explicitly stated details, two for making inferences, two for summarising the main idea, and two for recognising attitudes and emotions. Each item should have ABCD four options and the reading skill tested in each question should be marked." (Adapted from Lin & Chen 2024)

2.6.3.4 Automated assessment of student-written texts

Several studies used prompts to automate assessment (e.g., Cohn et al., 2024; Lee et al., 2024; Yancey et al.). With the exception of Pinto et al. (2023), these studies drew on specific rubrics and evaluation criteria to create prompts. For instance, Cohn et al. (2024) designed a prompt to score students' writing according to established criteria of conceptual knowledge and critical thinking (Bloom et al., 1971). While some studies used the established rubrics for automated essay scoring, such as IELTS Task 2 Writing (Mizumoto & Eguchi, 2023), TOEFL iBT Writing Rubric (Wang & Gayed, 2024) and CEFR Rubric (Yancey et al.), other studies designed prompts based on the customised rubrics (Bucol & Sangkawong, 2024; Shin & Lee, 2024; Tang et al., 2024). Table 4 provides some examples.

Table 4: *Prompts for automated scoring*

Overall strategy	Prompt elements	Prompt example
Chain-of-Thought (CoT), Role-Play	Student response. Reference to a pre-developed rubric with associated scores. The rubric must be included in the prompt or it could be used to develop a meta prompt.	"You are a teacher whose job it is to score middle school student's short answers to a formative assessment question. Students are asked the following question [...] You are to score the responses based on the following rubric and scoring framework: [rubric is based on elements from Bloom's taxonomy] In each response, use quotations from the student's response as evidence, tying it back to the rubric, and providing a score and explanation; e.g., "The student says X. The rubric states Y. Based on the rubric, the student earned a score of Z". (Adapted from Cohn et al. 2024)
Input-output (few-shot and zero-shot), role-play, CoT.	Pre-developed rubric with various proficiency levels. Few-shot examples, e.g., a small number of student-written responses with human scores for the proficiency levels in the rubric. A CoT trigger, e.g., some examples of human scoring with human-written chain-of-thought explanations and the proficiency level categories.	"Please act as an impartial IELTS examiner and evaluate the quality of the response provided by an IELTS candidate to the writing task displayed below. Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, you must classify the response on a scale of 0 to 9. Refer to rubric and examples below for your ratings". (Adapted from Lee et al. 2024)

2.6.3.5 From user-oriented prompts to back-end prompting

An interesting trend across the reviewed studies was the growing interest in development-oriented prompts to create autonomous AI chatbots capable of performing language-related tasks (e.g., Dang et al., 2023; Han et al., 2023; Schick et al., 2022). We identified several studies where back-end prompts were devised as part of strategies to create customised LLM-based chatbots or platforms. Examples include:

- free-form or structured writing and revision (Dang et al., 2023; Han et al., 2023; Liu et al., 2023)
- grammatical error correction (Loem et al., 2023)
- supporting collaborative language learning (Schick et al., 2022)
- automated scoring (Shin & Lee, 2024).

Compared to the prompts described in the previous sections, which are relatively accessible for non-experts while still requiring a degree of familiarity with the terminology and conventions of language learning and assessment, prompts in this category are too complex to report in a simplified, synoptic format. Indeed, they mainly operate as meta-prompts in the background, unseen by end-users. This trend can be interpreted in two ways. On the one hand, it reflects the sophistication and diversification of the GenAI landscape in education, with a great deal of research seeking to develop custom "closed off" solutions that meet market demands, rather than empower end-users through open-ended interactions.

On the other hand, it marks a clear shift, already noted in the more critical literature on educational technology, towards increased automation and standardisation (Perrotta, 2024; Selwyn, 2019). According to this more critical work, AI is the latest chapter in a long story of educational reform driven by market-based ideals, which have introduced narrow notions of technology-enabled personalisation that encourage transactional and individualistic approaches to teaching and learning (Pelletier, 2024).

2.7 Making sense of the literature on prompting

The purpose of this review was to examine the use of prompts in the research literature on GenAI in language learning and assessment. The review provided an analysis of research on prompting strategies and, where possible, examples of how prompts can be developed. It is worth noting however that prompting is a means to an end. As such, the studies we reviewed were not about prompting *per se* but instead mobilised prompting for a very distinct challenge: the development of AI-assisted or fully automated systems for language learning and assessment. In this concluding discussion, we wish to provide a synthesis of what is emerging from this effort.

To begin with, the research literature suggests that AI systems – for now at least – appear to excel at quantitative evaluation and can achieve high agreement with human raters (e.g., Bucol & Sangkawong, 2024; Cohn et al., 2024; L. Wang et al., 2024). However, these systems still have significant limitations when it comes to qualitative evaluation. The main difference is the degree of assessment depth: human assessors provide holistic, context-aware feedback that incorporates student history and developmental needs, while AI systems tend to focus on linguistic and structural elements, often missing deeper contextual factors (Guo & Wang, 2023; L. Wang et al., 2024).

The reliability of automated assessment systems is also worth noting. The studies we reviewed show significant variation based on implementation factors and context. Automated scoring systems can achieve high statistical reliability under controlled conditions, particularly when using refined prompting strategies and clear assessment rubrics (Lee et al., 2024; Tang et al., 2024). However, this reliability is heavily dependent on technical parameters and, unsurprisingly at this point, on prompt design. The research reveals persistent challenges in maintaining robust performance across different assessment contexts, with automated systems struggling the most when evaluating stylistic elements and conventions. Additionally, systems show varying levels of reliability based on response length and complexity, with performance typically degrading as text length increases (L. Wang et al., 2024; Yancey et al.).

Finally, the fundamental question of the suitability of large language models for language assessment reveals several systematic limitations that warrant careful consideration. A noticeable pattern across the research we reviewed shows that these models exhibit specific weaknesses in distinguishing between levels of understanding, particularly when evaluating higher-order thinking skills. The evidence indicates that they often misinterpret surface-level features (such as the use of certain words or phrases) as indicators of deeper understanding, leading to potentially flawed assessments (Cohn et al., 2024; Lee et al., 2024).

Language and cultural biases also emerge as a significant concern. Studies have documented systematic bias patterns in LLM assessments, particularly affecting non-native English speakers. For instance, research shows reduced agreement between human and LLM ratings for speakers of certain languages, including Telugu, Bengali, and Mandarin Chinese (Yancey et al.). This raises serious questions about assessment fairness and equity.

The research also reveals a concerning pattern of leniency and inconsistency in certain assessment contexts, with models showing greater leniency in evaluating organisational and content aspects compared to language use and sometimes struggling with maintaining consistent assessment standards across different response types (Shin & Lee, 2024). This suggests that current LLM systems might not be sufficiently calibrated for high-stakes educational assessment.

All told, these emerging findings and the accompanying issues and limitations suggest a complementary rather than a competitive relationship between humans and AI. This supports a common-sensical – but still worth advocating – argument for caution and circumspection around the potential of AI in education, and society more broadly.

Finally, three specific implications can be extrapolated from this literature review, which pave the way for the qualitative fieldwork reported in the next section.

1. Firstly, conceptualising prompting as a communicative genre is crucial to address a limitation in current research which often reduces prompting to a purely technical practice. While technical studies of prompt engineering have provided important insights into effective machine instruction, they often overlook the complex dimension of human-AI communicative interaction enabled by GenAI. This lens therefore opens up new analytical possibilities for understanding prompting as an emergent literacy practice. At the same time, framing prompting in terms of literacy invites a rethinking of the educational responsibilities in language learning and test preparation.

2. Secondly, examining differential experiences between teachers and students would address a notable gap in current research, where not much attention has been paid to how different stakeholders develop distinct patterns of engagement with prompting. This implication seems particularly important given that students and teachers often display different technological adoption patterns in educational settings (Granić, 2022).

3. Thirdly, a stronger focus on the cultural and social dimensions of prompt knowledge is needed to examine how non-expert users develop practical understandings of complex systems through everyday use rather than formal training.

Together, these three implications pointed to the need for naturalistic investigation of how prompting practices emerge and evolve in real-life language learning settings, informing the methodological approach and the specific focus of our subsequent empirical work.

3. RQ2: How is prompting for GenAI being used in naturalistic conditions?

3.1 Introduction to the qualitative fieldwork

The main premise underpinning the fieldwork is that prompting represents a key aspect defining a novel form of communicative interaction between humans and artificial intelligence (Guzman & Lewis, 2020). Having adopted this conceptual stance, we approached fieldwork as an opportunity to investigate the nature of prompting in a language education context delivering IELTS preparation courses: a real-world language school grappling with the disruption wrought by GenAI in language education and assessment. Our aim was to understand how people immersed in that context 'make sense' of the emerging forms of human-computer communicative interaction on their own terms and according to their pre-existing worldviews (Garfinkel, 1996). Within this conceptual framing, we approached the fieldwork as a case study (Yin, 2009). Additional key influences include classic and recent ethnographies carried out in educational settings (Monahan, 2005; Selwyn et al., 2017; Sims, 2017).

The fieldwork lasted 11 months (from February 2024 to December 2024) and took place in an English Language School in a large city in Australia. We worked with a small group of 24 participants: 16 students and 8 members of the teaching staff. Participants were all volunteers. They were interviewed and involved in group-based data collection events. Details are as follows:

- individual interviews with staff (N=8) followed by a 'prompting workshop' on 16 May 2024 during which the same teachers experimented with prompts to support automated assessment
- individual interviews with students (N=16) followed by a prompting workshop on 4 July 2024 led by the research team; during this workshop, students experimented with prompts to support their English language skills and test preparation activities
- two classroom observations (8 August 2024 and 22 August 2024), during which the research team observed a teacher run 'prompt literacy' sessions with international students.

The fieldwork generated two types of qualitative data: transcribed audio (interviews but also naturally occurring talk recorded during workshops), and conversation excerpts with ChatGPT shared by participants. All activities involving prompting used the free version of ChatGPT. This choice was deliberate in order to reflect as much as possible authentic conditions of use, especially for students.

The study received ethical approval from the University of Melbourne[1] and all participants gave their consent to the use of their data for research purposes.

Table 5 reports some basic descriptive information about the participants[2].

[1] Review reference: 2024-28467-50433-4

[2] Disclosure: teacher #1 was a key informant during fieldwork as she held a formal role as innovation and technology specialist at the school. She acted as a key point of contact and a 'gatekeeper' for the research team, assisting with recruitment of additional participants and helping to arrange workshops on site.

Table 5: *Research participants*

	ID	Age	Gender identification	L1	Nationality
1	Student#1	20-25	Female	Japanese	Japan
2	Student#2	40-45	Male	Japanese	Japan
3	Student#3	35-40	Female	Spanish	Colombia
4	Student#4	25-30	Male	Arabic	Saudi Arabia
5	Student#5	35-40	Female	Spanish	Colombia
6	Student#6	35-40	Female	Mongolian	Mongolia
7	Student#7	35-40	Male	Japanese	Japan
8	Student#8	20-25	Female	Thai	Thailand
9	Student#9	30-35	Male	Arabic	Saudi Arabia
10	Student#10	25-30	Female	Dari	Afghanistan
11	Student#11	35-40	Female	Spanish	Colombia
12	Student#12	35-40	Male	Thai	Thailand
13	Student#13	25-30	Female	Mandarin	Taiwan
14	Student#14	25-30	Female	Arabic	Saudi Arabia
15	Student#15	25-30	Female	Japanese	Japan
16	Student#16	25-30	Female	Japanese	Japan
17	Teacher#1	30-35	Female	English	Australia
18	Teacher#2	45-50	Male	English	Australia
19	Teacher#3	60-65	Female	English	Australia
20	Teacher#4	50-55	Male	English	Australia
21	Teacher#5	30-35	Male	English	Australia
22	Teacher#6	30-35	Male	English	Australia
23	Teacher#7	40-45	Male	Portuguese	Brazil
24	Teacher#8	55-60	Male	English	Australia

3.2 Analytical framework and coding strategy

The analysis followed a systematic process of thematic analysis (Silverman, 2021) that bridged our initial concerns from the literature review with emerging insights from the fieldwork. The analysis began with four broad deductive lenses.

1. Prompting as a communicative genre: this lens was shaped by a critical reading of the literature examined in the previous section that helped us frame prompting as a novel form of communicative practice. This view moves beyond the instrumental and technicist definitions of prompting as machine instruction, instead positioning it as an emerging form of literacy that requires both technical and communicative competence.

2. Differential experiences between teachers and students: this lens focused on how experiences of prompting differ among the primary stakeholders of GenAI – educators and learners – both viewed as non-expert users who may develop distinct patterns of communicative engagement with AI tools.

3. The cultural and social nuances of prompt knowledge and use: with a particular focus on how non-expert users develop practical and 'commonsensical' understandings of such a complex practice.

4. **Contextual application of prompting strategies:** the fourth lens concerned the 'foundational' and specific prompting strategies identified during the scoping review. This lens was created to help us examine whether and how these strategies may translate into viable teaching and learning practices.

This deductive framework led to a flexible empirical protocol that informed semi-structured interviews and observations.

Lens 1: prompting as a communicative genre

- How would you describe your communication style with AI tools?
- Can you walk me through a typical interaction when you're creating prompts?
- How does communicating with AI differ from other forms of communication?

Lens 2: differential experiences

- How did you first learn about prompting?
- How do your prompting practices compare to what you observe from your peers (if student) and/or colleagues (if staff)?
- What challenges have you encountered in using prompts? What challenges have you observed in others?
- Can you describe any instances where somebody else's use of AI surprised you?

Lens 3: cultural and social dimensions

- How do you share prompting knowledge?
- How does your cultural and/or linguistic background influence your approach to AI?
- How do you think different cultural contexts affect AI use in language learning?

Lens 4: Contextual prompting strategies

- Note patterns in how participants construct prompts
- Document instances of prompt sharing or discussion
- Record variations in prompting strategies across different tasks
- Observe informal knowledge exchange about AI
- Note emotional responses to AI interactions
- Observe how cultural factors influence AI tool use

These lenses served as an initial coding scheme for the fieldwork, but the analytical process remained open to inductive insights that might challenge or extend these preliminary categories (Glaser & Strauss, 2017). Through iterative engagement with the qualitative corpus, we observed how the initial broad theoretical categories were manifesting and/or changing. As a result, the themes that emerged through this process reflect both our initial theoretical concerns grounded in a critical interpretation of the prompting literature, and the unexpected ways participants in a real context of language learning and test preparation are adapting to, resisting, and re-imagining their relationship with GenAI tools and prompting in particular.

The themes are examined in detail in the following sections. Each theme captures a different facet of how teachers and students are making sense of prompting for GenAI, from challenges relating to practical implementation to deeper questions relating to identity and emotions. Excerpts from interviews and observational notes are used for illustrative purposes. Interviews with international students have been slightly edited for clarity, comprehension and readability.

3.3 Findings

3.3.1 Describing the context

The research context is an English language school operating in a large coastal city in Australia and affiliated with its oldest and most prestigious university. In the words of Teacher#3 (60-65, female), who has been employed at the school for more than 20 years:

> Teacher#3 (60–65, female): [Our school] has had a number of iterations, but it's a very long-standing language school here in [city], I think it has a reputation for delivering courses that are well designed, with teachers that are skilled.

The school was established in 1986 and offers various pathways into Australian education, such as:

- a bridging program for students who have already received an offer to enrol into a course conditional upon meeting specific English language requirements
- direct entry pathways into a range of TAFE (Technical and Further Education) institutions
- high school pathways focused on academic language skills recognised by government and private schools in Australia.

The school has a long-standing relationship with IELTS, first as a testing centre and more recently, as a preparation school. They have experienced a reduction in student enrolments over the years as students have more options and avenues to undertake their IELTS prior to entering Australia. Most of these students are 'predominantly from Asian countries' (Teacher#1, female, 30–35).

> Teacher#1 (female, 30–35): The largest nationality group would be Thai, Chinese, Turkish, Vietnamese. A lot of Vietnamese students – European not so much. We have a small smattering of various nations, Latin America. We have a lot of students from Colombia, Chile, Brazil...but predominantly Asian countries.

Most of these students join one of the pathways seeking higher IELTS scores to meet specific entry requirements set by Australian institutions. The IELTS preparation course runs for 10 weeks and offers a program of daily classes focused on the four components of the test (Reading, Writing, Listening and Speaking). The course was described as rather labour-intensive for staff and students alike.

> Teacher#3 (female, 60–65): We use Ready for IELTS [researcher note: a widely adopted textbook], as well as some external resources as well. And we do have regular practice tests. Each week, students are taking a full reading and listening test and each week, we alternate formative and summative writing tasks. In one week, we might have a summative task and a formative task. The following week, it will be swapped, so quite a bit of marking for teachers, but students come into this course with an expectation that they will practice the IELTS test and get a lot of feedback. So, it is quite a dense curriculum, and we've managed to fit it all into 10 weeks.

What transpires from the quotes above is that teachers at the language school deal with heavy workloads in terms of assessment – mostly related to the writing component of the IELTS – and with international students' expectations of personalised communicative support tailored around variable levels of English proficiency. With this contextual information as background, we proceeded to investigate experiences around GenAI and prompting amongst teachers and students.

3.3.2 The teacher perspective: how are teachers delivering IELTS preparation courses engaging with, and making sense of, GenAI prompting?

3.3.2.1 Emergent pedagogical integration – 'AI cheating' encounters shaped subsequent pedagogical uses

When asked about their experiences with GenAI and prompting, only two teachers described themselves as intermediate users, and both only ever used the free version of ChatGPT: Teacher#1 (female, 30–35) – our gatekeeper and 'innovation champion' (as previously stated, this was a key informant during fieldwork and held a formal role as innovation and technology specialist at the school) – and Teacher#5 (male, 30–35). All other teachers described themselves as familiar with the technology and/or possessing a superficial understanding of prompting methods and best practices, but keen to learn and become more proficient.

Our initial interactions with teachers revealed that the first meaningful encounters with AI technology at the school came not through agentic exploration, but reactively through institutional plagiarism detection systems. The experience was complicated by the unreliability of these systems, creating additional challenges for teachers attempting to maintain integrity standards in their classes. To be more precise, all participants stated that their very first experience with GenAI were reports of student misuse from Turnitin. Turnitin is a widely adopted platform that uses algorithms trained on vast repositories of student coursework to check for plagiarism. Following the global surge of GenAI use among students, Turnitin quickly deployed an AI detection tool in April 2023. The tool was subsequently found wanting and prone to mistakes and false positives.

> Teacher#8 (male, 55–60): I became aware of it in the last year of students using it, and in a negative way, with students using it to cheat in online exams or, and with the Turnitin AI detection tool it became obvious. I haven't actually used it to make materials or anything yet.

> Teacher#2 (male, 45–50): The Turnitin detection came in suddenly but was discredited almost as quickly. I don't know if it's regained its credit in terms of the validity of that score.

> Teacher#1 (female, 30–35): I've had colleagues come to me [with a Turnitin score] and say 'this paper has 60% AI'. I've spoken to those students, and they insist that they haven't used it, and you look at other students' work, which is in terms of the language at a very similar high level, and that might return a zero. So it's very inconsistent, it appears anyway. And it's just unclear how much we can rely on that score.

This initial exposure to the more 'problematic' side of AI use shaped teachers' subsequent engagement and led to a rather pragmatic attitude where the initial goal was not to seek efficiencies or professional empowerment, but to minimise student misuse. The general approach was therefore to model some basic 'safe' prompting strategies for test preparation that students could then replicate in their own time, thus maintaining a degree of pedagogical control over students' tendency to offload key learning tasks (especially writing) onto GenAI.

> Teacher#6 (male, 30–35): I'm not using prompts to tell a student 'You're at level two as a speaker' or whatever, but if you want to create a gap-fill exercise it's pretty good. I'll ask it [ChatGPT] to give me the answer and write down the answers and maybe write a table with the mistake in one column and the correction in another column. The correction can be a bit off sometimes. Then it's up to the student to model the prompts later for themselves.

These initial experiments with 'pedagogical prompting' also highlighted some interesting emotional responses to GenAI, and in particular, a certain amount of discomfort towards the overly obsequious tone of GenAI chatbots, which was viewed as a threat to the authenticity of the learning experience. This point emerged during the prompting workshop with teachers and sparked a discussion about the pedagogical value of challenging students in a more direct fashion, highlighting the importance of maintaining a degree of 'productive tension' during pedagogical interactions. In the absence of such a challenge, the deceptively smooth interactions with a generative agent might lead students to overconfidence, passivity, and to a false sense of progress in language acquisition. Teacher#6 observed that 'It's [ChatGPT] always super positive. It's always like, 'yeah, I can do that. Yeah, great', never gets angry. Like Scarlett Johansson in the movie, *Her*'. This encapsulates a crucial concern about the authenticity of AI-mediated learning experiences. However, there was an awareness that this issue is caused and made worse by unsophisticated and overly direct prompts and can be mitigated through better and more pedagogically informed instructions, pointing to the importance of developing more sophisticated prompting strategies aligned with established learning theories. In fact, the teachers' discussion revealed an emerging understanding of how prompt engineering could be aligned with established educational frameworks:

> Teacher#8 (male, 55–60): You could prompt it to do scaffolding, you know, like your zone of proximal development and keep pushing you slightly beyond your limits?
>
> Teacher#5 (male, 30–35): It would all come down to the prompt like, maybe you think I want to be pushed a little bit above my limit, then I could you could put it in the prompt and then it could prompt and say: I want you to be a certain character.

The recognition of the potential to integrate AI-assisted instruction with Vygotskian principles of scaffolding (Vygotsky, 1978) suggests a pathway toward more sophisticated applications of AI in language teaching. It is worth reminding that these principles, which are commonplace in educational theory and practice, refer to communicative and social assistance provided by a 'more experienced other' to a child or less-experienced peer. Such assistance 'scaffolds' the transition from a developmental stage to the next, which occurs when the child or less-experienced peer can complete a task without assistance. There is some interest in the use of AI – and GenAI in particular – to support scaffolding, but the methodological and empirical merit of these early efforts remains unclear (Cai et al., 2024).

3.3.2.2 Cultural and contextual dynamics – teachers' folk theories of student AI use

This theme explores how teachers develop informal theories about student engagement with AI, revealing the interaction between perceived motivations and cultural factors. The notion of folk theory originated in ethnomethodology and anthropology (Gelman & Legare, 2011; Ytre-Arne & Moe, 2021), and it refers to mostly 'heuristic' or experienced-based and intuitive strategies employed by people under deeply contextual circumstances that are not formal or based on scientific reasoning. These theories play a fundamental role in processes of sensemaking, suggesting that pedagogical uses of AI are shaped through daily observations and interactions as much as through formal theoretical frameworks or institutional policies.

A key aspect of these folk theories centres on the relationship between students' economic circumstances and their approach to AI use. Teachers observed distinct patterns between scholarship-supported students and self-funded students, suggesting that financial investment in education significantly influences attitudes toward AI adoption (as noted by Teacher#7 below).

> Teacher#7 (male, 40–45): Our course here is an expensive thing, right? It's not a cheap course, and depending on where you come from, it could be cheap or it could be expensive. Some of our students are funded by their government for a scholarship so they are much keener to work hard, because they received the chance to study for free at a university.

Teachers' observations also revealed a sophisticated understanding of how AI might be reshaping the broader motivational spectrum of the language education marketplace, with students engaging with AI to gain an advantage during exams and tests but still pursuing the valuable experience of travelling and learning language through immersion in a foreign cultural context.

> Teacher#5 (male, 30–35): Yeah, you might just notice that the market itself gets bigger because more people are able to access language education through free services. Students might start on AI and then they might, you know, they might say 'I'm actually pretty good at English, now...let's go to an English school and tweak it a bit, so I can pass my IELTS test or something like that'. Predominantly, students come to study English in Australia to, to like, to have that cultural experience.

Particularly noteworthy is teachers' recognition of how peer influence and fear of missing out (FOMO) drive AI adoption. Their observations suggest that AI use isn't necessarily correlated with academic struggle, but rather with students' perceived competitive advantage granted by this new and seemingly transformative technology:

> Teacher#2 (male, 45–50): Students who are cheating and using AI aren't necessarily the weaker students, and that's interesting, yeah...it's an issue of perception. I think that students enter a class with other students and hear about AI and perhaps feel 'I'm going to be at disadvantage unless I engage with this technology.'

To summarise, these folk theories serve two important functions. Firstly, they help teachers make sense of rapidly changing classroom dynamics in the absence of established research or institutional guidelines. Secondly, they illuminate how teachers actively negotiate their own relationship with AI through an approximate and emergent understanding of students' experiences and practices.

3.3.3 The student perspective: how are students enrolled in IELTS classes engaging with, and making sense, of GenAI prompting?

3.3.3.1 Lived experience as a motivator of language education and AI adoption

This theme is concerned with the motivations that underpin engagement with generative AI and prompting among language students preparing for their IELTS. The theme somewhat confirms the 'folk theories' we observed among teachers (see previous section) showing how lived experiences of economic and educational migration shape students' worldviews around language education, and subsequently around AI use for test preparation.

> Student#4 (Male, Arabic L1, 25–30): I am from Saudi Arabia. It's a country in the Middle East. I'm a paramedic. I did my Bachelor's and graduated in 2018 back in my country. Since then, I have been working as a paramedic and recently I had an opportunity to do a master's degree. I got a scholarship from my company, and they support me by paying some of my living expenses. As a paramedic, I studied all my subjects in English. And it's a part of my scholarship to improve my English before I start my master's degree. So, the plan is I will go to [Australian university] next March. Till this time, I need to improve my English. I need to get an IELTS test.

Student #2 (male, Japanese L1, 40–45): I wanted to be accepted into [university name], in the architecture department, because originally I'm an architect in Japan, and working as an architect in my city, but I just only did my Bachelor's in a Japanese university. I wanted to continue my academic career in Australia and get a good IELTS score.

Student #10 (female, Dari Persian L1, 25–30): Sometimes I think I'm the lucky one out of millions of Afghani girls. Because, you know, after the Taliban and all other things in my country, education is hard for girls. That is why, when I got, like, the scholarship from [university name], I didn't believe it!

Student#3 (female, Spanish L1, 35–40): I am from Colombia. My main reason for learning English is because it's required by my job. I used to work for an international company and the main language that we used was English. I had a senior job but I felt that I needed to improve my English. I studied English at school and university, but sometimes I find it difficult to speak and understand when I listen, so I decided to come here to learn English. I also wanted to have an experience of living in another country, very different from my country. The experience has been good in some ways. This school is really good and the teachers support you. But in some other aspects, my experience in Australia has not been like I was hoping, because I have developed some health problems that maybe don't allow me to be focused on learning. Everyone has a different process to learn a new language...

Under these varied and sometimes challenging circumstances, turning to GenAI is often a reluctant choice that goes against individual learning preferences. For many, it is the result of real or perceived peer influences, fuelled by a vague fear of being 'left behind' while others might be getting ahead in the competitive and time-pressured arena of economic and educational mobility. The fact that these opportunities are ill-defined and difficult to articulate suggests an emotional attitude towards GenAI, rather than a cognitive and instrumental one animated by rational goals and clear test preparation strategies.

Student#6 (female, Mongolian L1, 35–40): Everybody is talking about this stuff... 'You should use this or that', so I thought maybe I should learn. Maybe it's going to help me with my assessments! Gosh, I struggle a lot with the Writing Task. They told me not to worry, that I should try it – I know people who use it all the time for their assessments, but I worry it's bad for your learning...but I think I might use it in the future! [laughs]

Student#4 (Male, Arabic L1, 25–30): These tools [GenAI chatbots] are helpful, they are very helpful. They can make you improve in many ways. I believe they are useful... [researcher note: interviewee is invited to elaborate but struggles]. They...they can give you different ways of saying any phrase, any sentence, and help you change the sentence structure. But in my case, I believe that I'm not very used to them. I prefer the old-fashioned way of learning, because I'm used to this way.

Student#3 (female, Spanish L1, 35–40): I use [chatbots] to correct my essays and try to understand what my mistakes are...but I prefer books. I found out that's how I prefer to learn English. I use this technology because we are surrounded by technology. We need it, yeah.

Such emotional engagement with AI is interesting because it contradicts research and mainstream commentary around GenAI and prompting, where themes of efficiency, rational choice, and cognitive empowerment through offloading of secondary tasks are prominent.

This finding supports this report's thesis that prompting should not – or not only – be considered as a technical skill to instruct machines, but as a form of socio-cultural communication. This communication is now occurring between people and technology through codified natural language, and it remains profoundly social, that is, influenced by cultural and affective assumptions where the real, but more often imagined, potentials of this novel and opaque technology are negotiated: sometimes contested, other times uncritically accepted at face value.

3.3.3.2 Tactical prompting

This theme relates specifically to the range of, and proficiency with, prompting that we observed amongst our student participants. The first noteworthy insight is that, when students were asked to share their experiences with prompting in the context of IELTS preparation, it became immediately clear that these experiences were limited to seeking support around the Writing component of the test. This is perhaps unsurprising, given that writing remains for the time being the most immediate and accessible mode to interface with a chatbot, but of course, it is not indicative of future scenarios as voice-based interaction is poised to be increasingly integrated in GenAI systems.

Consistent with the emotional nature of student engagement with GenAI noted previously, this theme highlights that participants were ill-equipped to make an informed and autonomous use of sophisticated prompting strategies, such as those that emerged from the literature review. Nonetheless, we observed several attempts to use prompts. These attempts reflect a tension between an unsubstantiated belief in the 'promise' of GenAI as a powerful, almost magical, assistive tool, and the reality of prompting as a complex linguistic and cognitive practice that requires logical reasoning skills and a time-consuming inclination towards conceptual decomposition.

This tension informed a diverse range of low-level prompts which were perhaps linguistically and logically unsophisticated but were also far from naïve. In fact, these prompts reflected pragmatic and instrumental goals, with students acutely aware of the AI's ability to reproduce a 'good enough' approximation of personalised IELTS feedback. We asked participants to share excerpts of these tactical conversations with ChatGPT and categorised the prompts in Language Learning Prompts (Non-IELTS Specific) and IELTS-Specific Prompts. Some examples are reported below.

Language Learning Prompts (Non-IELTS Specific): in this category we noted a pattern whereby students attempted a progression from simple to more complex prompts as they refined their queries.

- Basic: 'Please tell me the way to practice my English skills.'
- Time-constrained: 'Please tell me the way to practice my English skills more in only a week.'
- Specific schedule request: 'Please suggest me a schedule to practice my English skills more in only a week, if I have only five hours per day.'
- Vocabulary-specific: 'I'd like to increase my vocabulary especially in phrasal verbs, how can you help?'

IELTS-Specific Prompts: like the generic language learning prompts, these IELTS-specific prompts (some examples below) were also cautiously progressive, but progression was not underpinned by a clear task-decomposition strategy and was instead recursive (going in circles), eventually regressing towards direct requests for sample text ('give me some examples').

- Direct request for writing questions: 'Could you give me writing questions for the IELTS exam?'

- Assessment requests: 'You are an IELTS instructor. Can you give some feedback on grammar mistakes in my writing?'
- Specific scoring requests: 'According to the IELTS criteria band 7 can you score my previous writing?'
- Grammar-specific queries: 'When can I use zero articles in my writing?'
- Spanish language request (showing multilingual support seeking): 'Practicar speaking de IELTS.'
- General IELTS: Basic guidance seeking: 'How can I learn for IELTS exam?'

Overall, we noted that students possessed sufficient commonsensical knowledge of prompting to understand they needed to include specific linguistic criteria and/or IELTS terminology in their prompts to increase the relevance and precision of feedback.

> Student #8 (female, Thai L1, 20–25): The IELTS people – they want cohesion and grammatical accuracy, and also stuff like introductions, topic sentences, inferences and conclusions.

> Student#12 (female, Thai L1, 35–40): I use a prompt where I ask ChatGPT to check my essay as an IELTS instructor in terms of the IELTS criteria such as good structure, grammatical accuracy, using appropriate vocabulary and using good cohesion.

> Student#5 (Female, Spanish L1, 35–40): In my prompts I ask it to check the grammatical structure, correct mistakes of my essays, and find more accurate and formal words, and check coherence and cohesion.

Despite such tactical and situational approach to prompting, which runs counter to the more strategic methods found in the prompt engineering literature, students remained alert to GenAI's shortcomings and to the fact that while its general-purpose nature may be a source of useful feedback, it remains less effective than more specialised AI-based technologies.

> Student#6 (female, Mongolian L1, 35–40): When I was in Mongolia, I used to get feedback from my teacher and pay like $30 an hour, because grammar has always been a problem for me and it still is now. Now with ChatGPT I can ask it to be an IELTS instructor and ask 'Can you give me some feedback on my grammar?' [laughs] but actually I don't want to use ChatGPT to write correctly, because I already use Grammarly which is easier and I can use while I write! It's easier, why should I learn a new application?

The trustworthiness of GenAI's feedback was also an issue during our discussions with students. Consistent with the tactical nature of prompting described earlier, we noted that AI's trustworthiness was also negotiated as a contingent and local dynamic, that is, unfolding in the situational immediacy of the language school, where teachers remained key figures of authority. The next quote is quite telling as it implies that ChatGPT 'misunderstood' or provided an erroneous evaluation, not because of its structural and reasoning limitations, but simply because it did not align with the assessment the teacher provided that very same morning.

> Student#16 (female, Japanese L1, 25–30): I ask ChatGPT to give me feedback on my writing from an IELTS perspective, assessing cohesion, grammatical accuracy, academic vocabulary and it always responds by saying 'your essay is generally cohesive, but there are a few areas where the flow can be improved, or something like that'...To be honest, I wasn't happy with it because it misunderstood. My teacher revised it and he told me that it was not appropriate, but ChatGPT said it was!
>
> Researcher: So you trust your teacher more than ChatGPT?
>
> Student#16: Yeah! [laughs]

This confirms the ethnomethodological tenet according to which trust is, for most people at least, a socially and situationally produced contract rather than an overarching and decontextualised philosophical principle (Boltanski & Thévenot, 2006; González-Martínez & Mlynář, 2019).

4. Conclusion

This study showed that GenAI prompting holds promise for assisting several language-related tasks and testing processes, including some directly relevant to the IELTS test, but weaknesses in terms of robustness, bias and consistency remain. Good prompting can mitigate these weaknesses, but its implementation requires a perceptual and epistemological shift – from a view that pursues the automated and fast production of outputs to one that acknowledges a need for the laborious and slow curation of AI's behaviour through communicative competence. Throughout this report, we repeatedly critiqued the simplistic view of prompting as a technique for machine instruction and often reiterated the value of adopting such a communicative lens. Our overarching position, supported by our systematic engagement with the literature and by our qualitative fieldwork, is that generic and decontextualised 'prompting frameworks' for language education and language test preparation are certainly possible, but they are arguably undesirable. Instead, a contextual and indeed communicative understanding of prompting is more likely to foster favourable pedagogical conditions leading to the development of prompt literacies. By encouraging literacies rather than techniques, all key stakeholders with decisional power and influence (from language education providers to language testing organisations) can work towards a more naturalistic and therefore, sustainable use of GenAI in language learning and test preparation.

With this overarching position as background, we can now articulate the main implication emerging from the two strands of work undertaken for this study (the systematic scoping review and the qualitative fieldwork): a rift is beginning to open between technical understandings of GenAI prompting on the one hand, and real-life communicative/pragmatic scenarios on the other. While prompts in the real world of language education and testing are developing along a communicative and pragmatic trajectory, the field of 'prompt engineering' keeps developing along a technical one, resembling a specialised area of expertise for the AI development community and for power users, rather than a democratic and accessible competence for the broader population.

Our engagement in a language school, during which we deliberately examined the lived experiences of teachers and students making sense of prompting – and of GenAI more broadly – revealed that emerging forms of human-computer communication are also shaped by informal 'folk theories' and by the rich background of subjective and cultural motivations that underpin learner choices in language learning and testing. This background cannot be fully understood without an informed appreciation for the complex and often idiosyncratic trajectories of educational and economic mobility. Our findings showed that, even when unacknowledged, these cultural and economic factors shape the informal theories held by teachers and students about AI. Indeed, our study's exploratory foray into communicative sensemaking suggests that GenAI prompting should be viewed as one part of a complex ecosystem of teaching and learning strategies, competing technological affordances, social pressures, and economic considerations that shape patterns of adoption.

Our understanding of prompting patterns among non-experts must therefore go beyond a simplistic binary of effective versus non-effective prompting, to incorporate many more nuanced gradations that reflect contextual and socially shaped communicative goals. Effective advice for educators and learners needs to account for these nuances rather than adopting a one-size-fits-all approach to GenAI prompting in language education and test preparation. For instance, naïve but progressive prompting should not be viewed as the result of a superficial engagement with an AI's powerful reasoning affordances, but as a pragmatic and tactical behaviour that has its own goals and rationalisations, and as such should not be dismissed outright or discouraged in favour of more supposedly 'sophisticated' forms of prompting. Instead, this tactical and situational prompting should be recognised and discussed openly through deliberate linguistic and cognitive mediation to assist the transition to more strategic prompting, while maintaining the contextual and subjective investment that makes naïve prompting more salient and meaningful for users.

We hope that the further study of these mediation processes will be undertaken in the future. As part of these future efforts, students should be guided in developing personalised communicative approaches with AI through thoughtfully-designed prompts that recognise multiple pragmatic and communicative scenarios: AI can serve as a supplementary resource when teachers are unavailable, while also creating simulative environments where specific skills and competencies can be practiced in a low-risk settings.

References

Aryadoust, V., Zakaria, A., & Jia, Y. (2024). Investigating the affordances of OpenAI's large language model in developing listening assessments. *Computers and Education: Artificial Intelligence*, 6, 100204. https://doi.org/10.1016/j.caeai.2024.100204

Bai, S., Gonda, D. E., & Hew, K. F. (2024). Write-Curate-Verify: A Case Study of Leveraging Generative AI for Scenario Writing in Scenario-Based Learning. *IEEE Transactions on Learning Technologies*, 17, 1313–1324. https://doi.org/10.1109/TLT.2024.3378306

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., . . . Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI Feedback*. arXiv. https://doi.org/10.48550/arXiv.2212.08073

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, 57. https://doi.org/10.1016/j.asw.2023.100745

Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024). Developing evaluative judgement for a time of generative artificial intelligence. *Assessment & Evaluation in Higher Education*, 49(6), 893–905. https://doi.org/10.1080/02602938.2024.2335321

Bezirhan, U., & Von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *Computers and Education: Artificial Intelligence*, 5, 100161. https://doi.org/10.1016/j.caeai.2023.100161

Bloom, B., Hastings, J., & Madaus, G. (1971). *Handbook on formative and summative evaluation of student learning*. McGraw-Hill.

Boltanski, L., & Thévenot, L. (2006). *On justification: Economies of worth*. Princeton University Press.

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*. arXiv. https://doi.org/10.48550/arXiv.2005.14165

Bucol, J. L., & Sangkawong, N. (2024). Exploring ChatGPT as a writing assessment tool. Innovations. *Education and Teaching International*, 1–16. https://doi.org/10.1080/14703297.2024.2363901

Cai, L., Msafiri, M. M., & Kangwa, D. (2024). Exploring the impact of integrating AI tools in higher education using the Zone of Proximal Development. *Education and Information Technologies*, 1–74.

Chan, K. W., Ali, F., Park, J., Sham, K. S. B., Tan, E. Y. T., Chong, F. W. C., Qian, K., & Sze, G. K. (2025). Automatic item generation in various STEM subjects using large language model prompting. *Computers and Education: Artificial Intelligence*, 8. https://doi.org/10.1016/j.caeai.2024.100344

Chen, X., Li, J., & Ye, Y. (2024). A feasibility study for the application of AI-generated conversations in pragmatic analysis. *Journal of Pragmatics*, 223, 14–30. https://doi.org/10.1016/j.pragma.2024.01.003

Cohn, C., Hutchins, N., Le, T., & Biswas, G. (2024). A Chain-of-Thought Prompting Approach with LLMs for Evaluating Students' Formative Assessment Responses in Science. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21), 23,182–23,190. https://doi.org/10.1609/aaai.v38i21.30364

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: the state of the field. *International Journal of Educational Technology in Higher Education*, 20(1), 22.

Dang, H., Goller, S., Lehmann, F., & Buschek, D. (2023). Choice over control: How users write with large language models using diegetic and non-diegetic prompting. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: insights into efficacy and ENL student preference. I*nternational Journal of Educational Technology in Higher Education*, 20(1). https://doi.org/10.1186/s41239-023-00425-2

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2024). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.13544

Garfinkel, H. (1996). Ethnomethodology's program. *Social Psychology Quarterly*, 59(1), 5–21.

Gelman, S. A., & Legare, C. H. (2011). Concepts and folk theories. *Annual Review of Anthropology*, 40(1), 379–398.

Glaser, B. G., & Strauss, A. L. (2017). Discovery of grounded theory: Strategies for qualitative research. Routledge. https://doi.org/10.4324/9780203793206

González-Martínez, E., & Mlynář, J. (2019). Practical trust. *Social Science Information*, 58(4), 608–630. https://doi.org/10.1177/0539018419890565

Granić, A. (2022). Educational Technology Adoption: A systematic review. *Education and Information Technologies*, 27(7), 9725–9744. https://doi.org/10.1007/s10639-022-10951-7

Guo, K., & Wang, D. (2023). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*, 29(7), 8435–8463. https://doi.org/10.1007/s10639-023-12146-0

Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A Human–Machine Communication research agenda. *New Media & Society*, 22(1), 70–86. https://doi.org/10.1177/1461444819858691

Han, J., Yoo, H., Kim, Y., Myung, J., Kim, M., Lim, H., Kim, J., Lee, T. Y., Hong, H., Ahn, S.-Y., & Oh, A. (2023). RECIPE: How to Integrate ChatGPT into EFL Writing Education. *Proceedings of the 10th ACM Conference on Learning @ Scale*,

Hubert, K. F., Awa, K. N., & Zabelina, D. L. (2024). The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports*, 14(1). https://doi.org/10.1038/s41598-024-53303-w

Jovic, M., & Mnasri, S. (2024). Evaluating AI-Generated Emails: A Comparative Efficiency Analysis. *World Journal of English Language*,14(2), 502. https://doi.org/10.5430/wjel.v14n2p502

Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, 6, 100225. https://doi.org/10.1016/j.caeai.2024.100225

Koltovskaia, S., Rahmati, P., & Saeli, H. (2024). Graduate students' use of ChatGPT for academic text revision: Behavioral, cognitive, and affective engagement. *Journal of Second Language Writing*, 65, 101130. https://doi.org/10.1016/j.jslw.2024.101130

Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., Zhou, X., Wang, E., & Dong, X. (2023). *Better zero-shot reasoning with role-play prompting*. arXiv. https://doi.org/10.48550/arXiv.2308.07702

Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open*, 6, 100174. https://doi.org/10.1016/j.caeo.2024.100174

Lee, G.-G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, 6, 100213. https://doi.org/10.1016/j.caeai.2024.100213

Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2023). Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in English education. *Education and Information Technologies*, 29(9), 11,483–11,515. https://doi.org/10.1007/s10639-023-12249-8

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., & Rocktäschel, T. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ*, 339, b2700. https://doi.org/10.1136/bmj.b2700

Lin, Z., & Chen, H. (2024). Investigating the capability of ChatGPT for generating multiple-choice reading comprehension items. *System*, 123, 103344. https://doi.org/10.1016/j.system.2024.103344

Liu, M., Zhang, L. J., & Biebricher, C. (2024). Investigating students' cognitive processes in generative AI-assisted digital multimodal composing and traditional writing. *Computers & Education*, 211, 104977. https://doi.org/10.1016/j.compedu.2023.104977

Liu, Z., Litman, D., Wang, E., Matsumura, L., & Correnti, R. (2023). Predicting the quality of revisions in argumentative writing. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*.

Loem, M., Kaneko, M., Takase, S., & Okazaki, N. (2023). Exploring effectiveness of GPT-3 in grammatical error correction: A study on performance and controllability in prompt-based methods. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023).*

Lottridge, S., Ormerod, C., & Patel, M. (2024). Redesigning automated scoring engines to include deep learning models. In M. D. Shermis & J. Wilson (Eds.), *The Routledge International Handbook of Automated Essay Evaluation*. Routledge.

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. https://doi.org/10.1016/j.caeai.2023.100199

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2), 100050. https://doi.org/https://doi.org/10.1016/j.rmal.2023.100050

Mizumoto, A., Shintani, N., Sasaki, M., & Teng, M. F. (2024). Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment. *Research Methods in Applied Linguistics*, 3(2), 100116. https://doi.org/10.1016/j.rmal.2024.100116

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of Internal Medicine*, 151(4), 264–269.

Monahan, T. (2005). *Globalization, Technological Change, and Public Education*. Routledge.

O'Sullivan, B. (2023). Reflections on the application and validation of technology in language testing. *Language Assessment Quarterly*, 20(4-5), 501–511. https://doi.org/10.1080/15434303.2023.2291486

OpenAI. (n.d). Prompt engineering. Retrieved 17 February 2025 from: https://platform.openai.com/docs/guides/prompt-engineering

Oppenlaender, J., Linder, R., & Silvennoinen, J. (2024). *Prompting AI art: An investigation into the creative skill of prompt engineering*. arXiv. https://doi.org/10.48550/arXiv.2303.13534

Pelletier, C. (2024). Against personalised learning. *International Journal of Artificial Intelligence in Education*, 34(1), 111–115.

Perrotta, C. (2024). *Plug-and-play Education: Knowledge and Learning in the Age of Platforms and Artificial Intelligence.* Taylor & Francis.

Pinto, G., Cardoso-Pereira, I., Monteiro, D., Lucena, D., Souza, A., & Gama, K. (2023). Large language models for education: Grading open-ended questions using ChatGPT. *SBES '23: Proceedings of the XXXVII Brazilian Symposium on Software Engineering.* https://doi.org/10.48550/arXiv.2307.16696

Rodrigues, L., Pereira, F. D., Cabral, L., Gašević, D., Ramalho, G., & Mello, R. F. (2024). Assessing the Quality of Automatic-generated Short Answers using GPT-4. *Computers and Education: Artificial Intelligence*, 100248.

Roy, T., Kumar, A., Raghuvanshi, D., Jain, S., Vignesh, G., Shinde, K., & Tondulkar, R. (2024). SciSpace copilot: Empowering researchers through intelligent reading assistance. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21), 23,826–23,828. https://doi.org/10.1609/aaai.v38i21.30578

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). *A systematic survey of prompt engineering in large language models: Techniques and applications*. arXiv. https://doi.org/10.48550/arXiv.2402.07927

Santomauro, D. F., Herrera, A. M. M., Shadid, J., Zheng, P., Ashbaugh, C., Pigott, D. M., Abbafati, C., Adolph, C., Amlag, J. O., Aravkin, A. Y., Bang-Jensen, B. L., Bertolacci, G. J., Bloom, S. S., Castellano, R., Castro, E., Chakrabarti, S., Chattopadhyay, J., Cogen, R. M., Collins, J. K., . . . Ferrari, A. J. (2021). Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *The Lancet*, 398(10312). https://doi.org/10.1016/S0140-6736(21)02143-7

Schick, T., Dwivedi-Yu, J., Jiang, Z., Petroni, F., Lewis, P., Izacard, G., You, Q., Nalmpantis, C., Grave, E., & Riedel, S. (2022). *PEER: A collaborative language model*. arXiv. https://doi.org/10.48550/arXiv.2208.11663

Selwyn, N. (2019). *Should robots replace teachers?: AI and the future of education*. Polity Press.

Selwyn, N., Nemorin, S., Bulfin, S., & Johnson, N. F. (2017). *Everyday schooling in the digital age: High school, high tech?* Routledge.

Shibuya, Y., Hamm, A., & Cerratto Pargman, T. (2022). Mapping HCI research methods for studying social media interaction: A systematic literature review. *Computers in Human Behavior*, 129, 107131. https://doi.org/10.1016/j.chb.2021.107131

Shin, D., & Lee, J. H. (2024). Exploratory study on the potential of ChatGPT as a rater of second language writing. *Education and Information Technologies*. https://doi.org/10.1007/s10639-024-12817-6

Silverman, D. (2021). *Doing Qualitative Research*, 6th edition. Sage Publications.

Sims, C. (2017). *Disruptive Fixation*. Princeton Studies in Culture and Technology Princeton University Press.

Singh, V. K., Singh, P., Karmakar, M., Leta, J., & Mayr, P. (2021). The Journal Coverage of Web of Science, Scopus and Dimensions: A comparative analysis. *Scientometrics*, 126(6), 5113–5142. https://doi.org/10.1007/s11192-021-03948-5

Tang, X., Chen, H., Lin, D., & Li, K. (2024). Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments. *Heliyon*, 10(14), e34262. https://doi.org/10.1016/j.heliyon.2024.e34262

Vygotsky, L. S. (1978). *Mind in society: Development of higher psychological processes.* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Cambridge: Harvard University Press. J

Wang, L., Chen, X., Wang, C., Xu, L., Shadiev, R., & Li, Y. (2024). ChatGPT's capabilities in providing feedback on undergraduate students' argumentation: A case study. *Thinking Skills and Creativity*, 51. https://doi.org/10.1016/j.tsc.2023.101440

Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R. K.-W., & Lim, E.-P. (2023). *Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models.* arXiv. http://arxiv.org/abs/2305.04091

Wang, Q., & Gayed, J. M. (2024). Effectiveness of large language models in automated evaluation of argumentative essays: finetuning vs. zero-shot prompting. *Computer Assisted Language Learning*, 1–29. https://doi.org/10.1080/09588221.2024.2371395

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2022). *Self-consistency improves chain of thought reasoning in language models*. arXiv. https://doi.org/10.48550/arXiv.2203.11171

Wang, X., Zhong, Y., Huang, C., & Huang, X. (2024). ChatPRCS: A personalized support system for English reading comprehension based on ChatGPT. *IEEE Transactions on Learning Technologies*, 17, 1762–1776. https://doi.org/10.1109/TLT.2024.3405747

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., & Hajishirzi, H. (2022). *Self-Instruct: Aligning language models with self-generated instructions*. arXiv. https://doi.org/10.48550/arXiv.2212.10560

Wang, Y., & Zhao, Y. (2024). *Metacognitive prompting improves understanding in large language models*. arXiv. https://doi.org/10.48550/arXiv.2308.05342

Wang, Z., Valdez, J., Basu Mallick, D., & Baraniuk, R. G. (2022). Towards Human-Like Educational Question Generation with Large Language Models. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, & V. Dimitrova (Eds.), *Artificial Intelligence in Education. AIED 2022. Lecture Notes in Computer Science* (Vol. 13,355, pp. 153–166). Springer International Publishing. https://doi.org/10.1007/978-3-031-11644-5_13

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). *Chain-of-thought prompting elicits reasoning in large language models*. arXiv. https://doi.org/10.48550/arXiv.2201.11903

Woo, D. J., Wang, D., Guo, K., & Susanto, H. (2024). Teaching EFL students to write with ChatGPT: Students' motivation to learn, cognitive load, and satisfaction with the learning process. *Education and Information Technologies*. https://doi.org/10.1007/s10639-024-12819-4

Xu, Q., & Li, P. (2023). Computational modeling of language learning in the era of generative artificial intelligence: A response to open peer commentaries. *Language Learning*, 73(S2). https://doi.org/10.1111/lang.12605

Yancey, K. P., Laflair, G., Verardi, A., & Burstein, J. (2023). Rating short L2 essays on the CEFR scale with GPT-4. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023).*

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Yin, B., & Hu, X. (2023). *Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond*. arXiv. https://doi.org/10.48550/arXiv.2304.13712

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). *Tree of thoughts: Deliberate problem solving with large language models*. arXiv. https://doi.org/10.48550/arXiv.2305.10601

Yin, R. K. (2009). *Case Study Research: Design and Methods*. SAGE Publications.

Ytre-Arne, B., & Moe, H. (2021). Folk theories of algorithms: Understanding digital irritation. *Media, Culture & Society*, 43(5), 807–824.

Zhang, Z., Wang, S., Yu, W., Xu, Y., Iter, D., Zeng, Q., Liu, Y., Zhu, C., & Jiang, M. (2023). *Auto-instruct: Automatic instruction generation and ranking for black-box language models*. arXiv. https://doi.org/10.48550/arXiv.2310.13127

Zhou, T., Cao, S., Zhou, S., Zhang, Y., & He, A. (2023). Chinese intermediate English learners outdid ChatGPT in deep cohesion: Evidence from English narrative writing. *System*, 118, 103141. https://doi.org/10.1016/j.system.2023.103141

Zhu, J., & Liu, W. (2020). A tale of two databases: the use of Web of Science and Scopus in academic papers. *Scientometrics*, 123(1), 321–335. https://doi.org/10.1007/s11192-020-03387-8

Appendix 1: Table of shortlisted articles for the scoping review

Items	Source	Year	Authors	Title	Source title
1	Scopus	2024	Aryadoust V.; Zakaria A.; Jia Y.	Investigating the affordances of OpenAI's large language model in developing listening assessments	*Computers and Education: Artificial Intelligence*
2	Scopus	2024	Bai S.; Gonda D.E.; Hew K.F.	Write-Curate-Verify: A Case Study of Leveraging Generative AI for Scenario Writing in Scenario-Based Learning	*IEEE Transactions on Learning Technologies*
3	Scopus	2024	Banihashem S.K.; Kerman N.T.; Noroozi O.; Moon J.; Drachsler H.	Feedback sources in essay writing: peer-generated or AI-generated feedback?	*International Journal of Educational Technology in Higher Education*
4	Scopus	2023	Bezirhan U.; von Davier M.	Automated reading passage generation with OpenAI's large language model	*Computers and Education: Artificial Intelligence*
5	Scopus	2024	Bucol J.L.; Sangkawong N.	Exploring ChatGPT as a writing assessment tool	*Innovations in Education and Teaching International*
6	Scopus	2024	Chen M.-R.A.	Metacognitive mastery: Transformative learning in EFL through a generative AI chatbot fueled by metalinguistic guidance	*Educational Technology and Society*
7	Scopus	2024	Chen X.; Li J.; Ye Y.	A feasibility study for the application of AI-generated conversations in pragmatic analysis	*Journal of Pragmatics*
8	Scopus	2024	Cohn C.; Hutchins N.; Le T.; Biswas G.	A Chain-of-Thought Prompting Approach with LLMs for Evaluating Students' Formative Assessment Responses in Science	*Proceedings of the AAAI Conference on Artificial Intelligence*
9	Scopus	2023	Dang H.; Goller S.; Lehmann F.; Buschek D.	Choice Over Control: How Users Write with Large Language Models using Diegetic and Non-Diegetic Prompting	*Conference on Human Factors in Computing Systems, Proceedings*
10	Scopus	2023	Escalante J.; Pack A.; Barrett A.	AI-generated feedback on writing: insights into efficacy and ENL student preference	*International Journal of Educational Technology in Higher Education*
11	Scopus	2023	Guo K.; Wang D.	To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing	*Education and Information Technologies*
12	Manual citation searching	2023	Han, J., Yoo, H., Kim, Y., Myung, J., Kim, M., Lim, H., Kim, J., Lee, T. Y., Hong, H., Ahn, S.-Y., & Oh, A.	RECIPE: How to integrate ChatGPT into EFL writing education.	*Proceedings of the 10th ACM Conference on Learning @ Scale*
13	Scopus	2024	Jovic M.; Mnasri S.	Evaluating AI-Generated Emails: A Comparative Efficiency Analysis	*World Journal of English Language*
14	Scopus	2024	Koltovskaia S.; Rahmati P.; Saeli H.	Graduate students' use of ChatGPT for academic text revision: Behavioral, cognitive, and affective engagement	*Journal of Second Language Writing*
15	Scopus	2024	Lee G.-G.; Latif E.; Wu X.; Liu N.; Zhai X.	Applying large language models and chain-of-thought for automatic scoring	*Computers and Education: Artificial Intelligence*
16	Scopus	2023	Lee U.; Jung H.; Jeon Y.; Sohn Y.; Hwang W.; Moon J.; Kim H.	Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in English education	*Education and Information Technologies*

Items	Source	Year	Authors	Title	Source title
17	Scopus	2024	Lin Z.; Chen H.	Investigating the capability of ChatGPT for generating multiple-choice reading comprehension items	*System*
18	Scopus	2024	Liu M.; Zhang L.J.; Biebricher C.	Investigating students' cognitive processes in generative AI-assisted digital multimodal composing and traditional writing	*Computers and Education*
19	Scopus	2023	Liu Z.; Litman D.; Wang E.; Matsumura L.; Correnti R.	Predicting the Quality of Revisions in Argumentative Writing	*Proceedings of the Annual Meeting of the Association for Computational Linguistics*
20	Scopus	2023	Loem M.; Kaneko M.; Takase S.; Okazaki N.	Exploring Effectiveness of GPT-3 in Grammatical Error Correction: A Study on Performance and Controllability in Prompt-Based Methods	*Proceedings of the Annual Meeting of the Association for Computational Linguistics*
21	Scopus	2024	Meyer J.; Jansen T.; Schiller R.; Liebenow L.W.; Steinbach M.; Horbach A.; Fleckenstein J.	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*
22	Scopus	2023	Mizumoto A.; Eguchi M.	Exploring the potential of using an AI language model for automated essay scoring	*Research Methods in Applied Linguistics*
23	Scopus	2024	Mizumoto A.; Shintani N.; Sasaki M.; Teng M.F.	Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment	*Research Methods in Applied Linguistics*
24	Scopus	2023	Pinto G.; Cardoso-Pereira I.; Monteiro D.; Lucena D.; Souza A.; Gama K.	Large Language Models for Education: Grading Open-Ended Questions Using ChatGPT	The study explored the use of LLMs (Chat GPT) in correcting open-ended questions and feedback. The open-ended questions were related to software training namely, web application cashing and performance testing
25	WOS	2024	Ranade, N; Saravia, M; Johri, A	Using rhetorical strategies to design prompts: a human-in-the-loop approach to make AI useful	*AI & SOCIETY*
26	Scopus	2022	Schick T.; Dwivedi-Yu J.; Jiang Z.; Petroni F.; Lewis P.; Izacard G.; You Q.; Nalmpantis C.; Grave E.; Riedel S.	PEER: A COLLABORATIVE LANGUAGE MODEL	*11th International Conference on Learning Representations, ICLR 2023*
27	Scopus	2024	Shin D.; Lee J.H.	Exploratory study on the potential of ChatGPT as a rater of second language writing	*Education and Information Technologies*
28	Scopus	2024	Tang X.; Chen H.; Lin D.; Li K.	Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments	*Heliyon*
29	Scopus	2024	Wang L.; Chen X.; Wang C.; Xu L.; Shadiev R.; Li Y.	ChatGPT's capabilities in providing feedback on undergraduate students' argumentation: A case study	*Thinking Skills and Creativity*
30	Scopus	2024	Wang M.; Wang M.; Xu X.; Yang L.; Cai D.; Yin M.	Unleashing ChatGPT's Power: A Case Study on Optimizing Information Retrieval in Flipped Classrooms via Prompt Engineering	*IEEE Transactions on Learning Technologies*

Items	Source	Year	Authors	Title	Source title
31	Scopus	2024	Wang Q.; Gayed J.M.	Effectiveness of large language models in automated evaluation of argumentative essays: finetuning vs. zero-shot prompting	*Computer Assisted Language Learning*
32	WOS	2024	Wang, X.Z; Zhong, Y.H.; Huang, C.Q.; Huang, X.D.	ChatPRCS: A Personalized Support System for English Reading Comprehension Based on ChatGPT	*IEEE Transactions on Learning Technologies*
33	WOS	2022	Wang, Z.C.; Valdez, J.; Mallick, D.B.; Baraniuk, R.G.	Towards Human-Like Educational Question Generation with Large Language Models	*Artificial Intelligence in Education, Pt I*
34	Scopus	2024	Woo D.J.; Wang D.; Guo K.; Susanto H.	Teaching EFL students to write with ChatGPT: Students' motivation to learn, cognitive load, and satisfaction with the learning process	*Education and Information Technologies*
35	Scopus	2024	Yamashita T.	An application of many-facet Rasch measurement to evaluate automated essay scoring: A case of ChatGPT-4.0	*Research Methods in Applied Linguistics*
36	Scopus	2023	Yancey K.P.; LaFlair G.T.; Verardi A.R.; Burstein J.	Rating Short L2 Essays on the CEFR Scale with GPT-4	*Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pp. 576–584, July 13, 2023 c 2023 Association for Computational Linguistics*
37	Scopus	2023	Zhou T.; Cao S.; Zhou S.; Zhang Y.; He A.	Chinese intermediate English learners outdid ChatGPT in deep cohesion: Evidence from English narrative writing	*System*