Innovating Reading Materials Development Based on Generative AI Within the Backward Design Framework

Zhiqing Lin [1], Tingting Cui [2]
1 School of Foreign Languages, Guangdong Polytechnic Normal University, Guangzhou, China.
2 Center for linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, China.

Author Note
The authors declare no conflicts of interest to disclose. Correspondence concerning this article should be addressed to Tingting Cui, Guangdong University of Foreign Studies, Guangzhou, Guangdong, 510420, China.
Email: tingtingcui2024@163.com

Abstract

Tailoring reading textbooks for personalized learning is essential. However, this process is time-consuming, labor-intensive, and costly for language teachers. This study fills this gap using generative AI to develop customized reading materials based on the backward design framework, which can assess five reading subskills in the teaching objectives. 288 EFL young learners were invited to respond to the reading materials generated by generative AI within 45 minutes. The teaching materials developed by AI were evaluated within the backward design framework. The result found that generative AI could develop customized reading materials for the desired reading subskills. Psychometric models revealed that the difficulties of reading materials were adaptive to students' reading ability holistically and that the reading materials were unbiased toward different genders. Besides, English teachers believed that AI-generated reading materials were useful and helpful to language teaching and learning. This study provides a new method for developing personalized reading materials for language teaching, learning, and testing using generative AI, which has the potential to increase the usefulness and diversity of teaching materials efficiently.

Keywords: Reading materials, generative AI, backward design framework, reading subskills, language teaching

Received 8 January 2026; revised 13 February 2026; accepted 6 March 2026; available online 25 March 2026; Version of Record 30 March 2026.

Citation: Lin, Z. Q. & Cui, T. T. (2026). Innovating reading materials development based on generative AI within the backward design framework. *Journal of Language*, *2*(1), 38-72. https://doi.org/10.64699/26GXPJ1206

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

1. Introduction

Although reading textbooks are widely used by English teachers and students, they have several shortcomings that need to be addressed (Curiel & Durán, 2021). For instance, English teachers might feel that the materials in reading textbooks are not customized enough to meet students' needs since the exercises in the reading textbooks are sometimes oversimplified (Bouckaert, 2015). This is an urgent need to customize reading textbooks to provide more personalized reading materials for students (Gilmore, 2007). Besides, materials development for language learners is a time-consuming and costly task for language teachers, and English teachers often get stuck on the preparation and adaptation of materials in language teaching (Rathert & Cabaroğlu, 2021).

Therefore, this study aims to address these challenges by using generative AI (GPT-4o) to develop reading materials based on the reading textbooks automatically. The purposes of this study are threefold: (1) developing customized reading materials that can measure five desired reading subskills in the teaching objectives, (2) examining the acceptability of reading materials developed by AI with psychometric models, and (3) investigating teachers' acceptance of using generative AI in developing reading materials.

The significance of this paper lies in four aspects: First, it can empower English teachers to develop customized instructional reading materials, providing customized reading materials for students to develop their different reading subskills. Second, it offers a solution for generating personalized learning materials tailored to students' needs. Third, it enhances the efficiency of English teachers in developing customized reading materials for students,

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

alleviating their burdens in materials preparation. Fourth, it can provide high-quality quizzes for English teachers to check students' understanding of the reading courses and provide diagnostic feedback to students.

2. Literature Review

2.1 Challenges of Reading Textbooks in Language Teaching

Reading textbooks play an essential role in language teaching, which are also important learning materials for EFL students (Wang & Ping, 2023). However, studies have found that there is a mismatch between textbook designers and end-users (Tsagari & Sifakis, 2014). English teachers might encounter challenges when they endeavor to use ELT reading textbooks in the classroom for local needs (Merino & Metila, 2024).

The first challenge language teachers face is the lack of customized reading tasks that can cover different reading subskills. Traditional reading materials often rely on comprehension questions that involve superficial reading, limiting students' active engagement and deep learning (Freeman, 2014). The second challenge is that English teachers have limited time and experience in developing customized reading materials for instructional purposes since materials preparation is time-consuming and challenging (Norton & Buchanan, 2022).

These two problems pose great challenges for language teachers in the practice of teaching reading in the classroom and compromise the teaching effect of reading courses (Tomlinson, 2012). However, as far as the author's knowledge, no research has been conducted to address these challenges English teachers are facing. This underscores the importance of conducting empirical studies that can provide possible solutions to address or mitigate these

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

challenges, which serve as the rationale for conducting this research.

2.2 Reading Subskills in Teaching Reading Comprehension

Reviewing the reading subskills in language teaching is essential since it can help us better integrate educational technologies with the teaching objectives in reading textbooks. During the past two decades, different scholars have proposed different taxonomies for the reading subskills in language teaching (Aryadoust, 2020).

The classification of reading subskills might be different, but one consensus they share is that reading subskills can be divided into lower-order and higher-order subskills, covering different aspects of the reading process (Grabe & Stoller, 2011). This study focused on five reading subskills: (A1) “understanding the meaning of new words”, (A2) “understanding explicitly stated information”, (A3) “understanding implicitly stated information”, (A4) “recognizing the author’s attitude”, and (A5) “summarizing the main idea of the passage”.

These five subskills were chosen for the following reasons. The first reason is that these five reading subskills can cover both lower-order reading subskills and higher-order reading subskills, as the first and second subskills are lower-order subskills, and the remaining three subskills are higher-order reading subskills (Ilc & Stopar, 2014). Therefore, these five subskills can cover both lower-order and higher-order reading processes. The second reason is that these five reading subskills are also important subskills in language teaching (Bamford & Day, 1998).

However, the existing materials in the reading textbooks cannot provide customized reading materials for these important reading subskills, which is essential for the development of different reading subskills. Besides, developing such customized reading materials can also

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

lay the foundation for diagnostic reading assessment (Kim, 2014), which can help English teachers better understand students' strengths and weaknesses in reading subskills. Nonetheless, developing such customized reading materials for students is extremely time-consuming and costly, which presents great challenges for English teachers (Li & Gao, 2025). This underscores the importance and necessity of enhancing the efficiency in developing customized reading materials for students and teachers.

2.3 Application of Large Language Models in Materials Development

With the development of generative AI, it is possible to leverage large language models (LLMs) to address the challenges reviewed and discussed in section 2.1 (Denny et al., 2023) and develop customized reading materials for different reading subskills reviewed in section 2.2, empowering the generation of personalized materials for students.

Studies have demonstrated the potential of LLMs to generate human-like text in educational contexts (Jeon & Lee, 2023; Ahmed et al., 2024) and increase students' reading motivation (Yılmaz & Aydın, 2025). Nevertheless, the new topic of AI-assisted materials development has rarely been explored, with several exceptional cases. For example, Jin and Lu (2018) investigated text adaptation using a data-driven method. Besides, Chen & Wu (2024) endeavored to explore the potential of applying generative AI in materials development for teaching poetry. Another typical study was conducted by Xin (2024), who explored teacher perceptions of applying ChatGPT for reading materials development.

These studies have made important contributions to AI-assisted materials development and provide crucial directions for future studies. However, the limitation of previous studies

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

mentioned above is that none of them has paid attention to the materials development for different reading subskills. Therefore, how to operationalize the reading subskills in the teaching objectives in the process of AI-assisted materials development remains insufficiently explored. Moreover, whether the materials developed by AI are biased remains underexplored, which suggests that the acceptability of the materials remains unknown. Additionally, the acceptance of English teachers at middle school is inadequately studied. These gaps warrant the need to conduct new empirical studies on the feasibility, validity, and acceptance of AI-assisted materials development.

2.4 The Theory Framework for AI-assisted Materials Development

The AI-assisted materials development also entails a theoretical framework. The backward design framework proposed by Wiggins and McTighe (2005) was used in this study to guide the development and validation of the materials based on AI. This framework consists of three key elements, including (1) *Identify desired results,* (2) *Determine acceptable evidence,* (3) *Plan learning experiences and instructions*. "Identify desired results" focused on determining the targeted objectives of the materials. The theoretical definition of this stage dwells on checking whether the materials can measure the desired subskills. "Determine acceptable evidence" is a process in which material developers evaluate the extent to which the materials are acceptable for students. The theoretical definition of acceptable can be investigated from the perspective of validity, reliability, and fairness (Wiggins & McTighe, 2005). "Plan learning experiences and instructions" pertains to the process of developing teaching materials. The theoretical definition of this stage mainly involves teacher's experience

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

(acceptance) in developing materials (Richards, 2013).

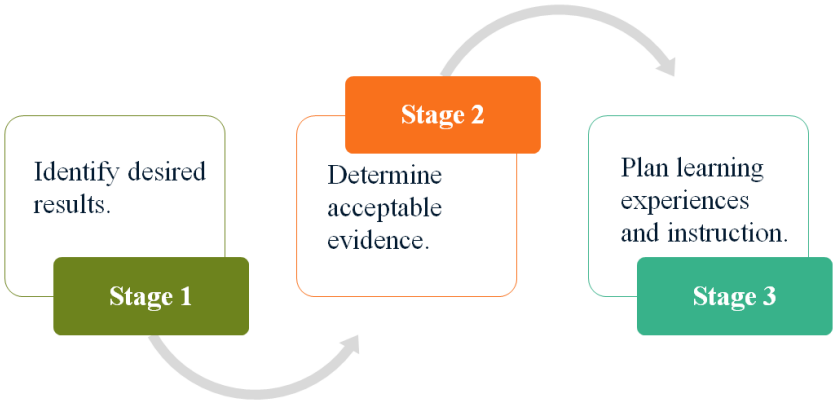

Figure 1. Backward design framework by Wiggins and McTighe (2005)

This framework was used in this study for the following reasons. The first reason is that this framework is widely used in materials development in language teaching. The second reason is that this framework can provide a roadmap to achieve student-centered materials design (Konttinen, 2022). The third reason is that this framework can provide systematic standards for the evaluation of the materials generated by AI in this study (Wiggins & McTighe, 2005).

2.5 Research Gap and Questions

This study aims to address the following gaps. First, most studies focused on general reading materials, but none of them have paid attention to developing personalized reading materials to cover both lower-order and higher-order reading subskills. Secondly, the acceptability of the reading materials developed by AI remains underexplored. Furthermore, there is a lack of research examining the development of teaching materials from the perspective of language teachers, particularly regarding their acceptance of using LLMs for AI-assisted materials development.

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

However, these gaps are the fundamental issues in technology-enhanced materials development. To address these gaps, this study aims to investigate the following three questions, which also correspond to the three stages of the backward design framework.

RQ1: To what extent can reading materials generated by generative AI assess desired reading subskills?

RQ2: To what extent are the reading materials developed by generative AI acceptable in terms of validity, reliability, and fairness based on psychometric models?

RQ3: To what extent do English teachers accept AI-assisted materials development according to their perceptions?

3. Methodology

3.1 Research Context

This research was conducted in the rural areas of southern China. The participants are in grade eight, and their average age is about 15 years old. We consulted the English teachers on the language proficiency of our participants. We were informed that their language proficiency was about A2 based on the CEFR standard, which provided insight into how to design and develop reading materials. Accordingly, we subsequently incorporated the expected difficulty of A2 in the prompt for LLMs when we designed the prompt for materials development.

3.2 Material Selection

The second step of materials development was about selecting the reading materials and the reading task. Three reading materials from an English textbook (Chen, 2013) in grade

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

eight were selected, and these passages have not been covered by the teacher in class yet. And one reading passage from grade nine was also selected. The reason is that, based on the theory of the zone of proximal development, it is appropriate to select the reading material that is somewhat beyond students' reading ability (Lantolf & Poehner, 2010).

The reason why we used the reading passages from the reading textbooks is that doing this can shed light on how teachers can use AI for materials development and lesson preparation. Besides, using the materials from textbooks can lower the chance of being biased. It is acknowledged that it is also possible to ask LLMs to develop the reading materials from scratch.

3.3 Materials Development

After selecting the proper reading passages from the English textbook, the next step was to develop reading materials via LLMs. In terms of the item type, multiple choice (MC) items were selected in that it is one of the most effective and commonly used item types to evaluate the student's learning effect. Moreover, MC questions were favored by teachers because this test format was easy to operate in class and could be used to promote classroom discussions and student engagement.

The materials entered into LLMs included the reading materials mentioned above and the prompt we designed. The prompt we designed was developed based on the principle of prompt engineering suggested by Zhou et al. (2022). Whenever the generated reading materials had some drawbacks, we also tried to improve the outcome using iterative prompting techniques (Lin, 2024; Shi et al., 2023). The key parts of the prompt we designed are listed below, and a more detailed prompt for reading materials development can be found in Appendix

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

A.

> "Now suppose you are an English teacher and your task is to develop reading materials for your students. You are going to develop some multiple-choice reading items for students during the English class. Their reading ability is around A2. Based on the passage I give you, you should generate five multiple-choice items with ABCD four options.
>
> One item for understanding the meaning of new words; one item for understanding explicitly stated information; one item for understanding implicitly stated information; one for recognizing the author's attitude and one for summarizing the main idea of the passage. "

We instructed GPT-4o to develop five questions for each of the four reading passages separately, and the five questions of each passage corresponded to these five reading subskills reviewed in section 2.2. Therefore, there were 20 MC items in total, which can be found in Appendix B.

Currently, users can have access to GPT-4o for a limited number of times every day. However, we feel this is not a problem since its alternative (GPT-4o mini) is open to every user without any constraints, and GPT-4o mini has a comparable capability in materials development compared with GPT-4o. Therefore, everyone can access this AI tool. Alternatively, users can also use other LLMs, such as Deepseek, which are freely open to the public.

Additionally, to enhance the replicability of this study, we also added a video-based tutorial for materials development, which can be found on our website for supplementary

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

materials (see the end of this paper). Hence, material developers and English teachers can use the prompt designed in this study to develop their reading materials based on other materials.

3.4 Material Administration

After generating the item based on the English textbook with LLMs, we administered the test formed by the reading materials generated by LLMs to participants. 288 participants received the paper-and-pencil test, and they were told to complete the task within 45 minutes independently. All test responses were collected and then digitized into the Excel sheet for further analysis.

3.5 Data Collection and Analysis

Table 1. Operational definitions of five reading subskills

Subskills	Definitions
A1	"Understanding the meaning of new words" involves utilizing the clues in reading contexts to infer the meaning of unfamiliar words (Scott & Nagy, 1997).
A2	"Understanding explicitly stated information" means identifying information directly presented without requiring inference or interpretation beyond the literal meaning (Fernández, 2008).
A3	"Understanding implicitly stated information" refers to grasping the underlying or hidden meaning not directly stated in the text (Lumley, 1993).
A4	"Recognizing the author's attitude" involves figuring out the writer's feelings, opinions, or perspectives toward the event (Aryadoust, 2020).
A5	"Summarizing the main idea" means synthesizing the main content or the theme of the whole passage (Li & Suen, 2012).

To investigate the first question, we invited five experts in applied linguistics to conduct expert judgment on the reading subskill each item can measure. The operational definition of

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

expert judgment can be found in Table 1. The background information of the five experts was related to language assessment and language teaching. Detailed information can be found in Appendix C. Whenever there were disagreements, experts negotiated with each other to arrive at final consensus.

To investigate the second question, we operationalized acceptability as validity, reliability, and fairness. We further operationalized the validity as the suitability of item difficulty based on item response theory (Lee-Ellis, 2009). The suitability of item difficulty can be investigated by an item-person map, which displays the item difficulty and student reading ability on the same scale. Therefore, we employed the Rasch model to estimate the item difficulty and student reading ability based on student responses collected during the experiment in the material administration. Then, we used the WrightMap package to plot the item-person map. We also used the Rasch model to analyze the reliability of the materials developed by AI. Besides, we operationalized and investigated the fairness by differential item functioning based on the Mantel-Haenszel test (Uiterwijk & Vallen, 2005) with the difR package (123 males and 165 females). We focused on the gender variable since gender is an important consideration in materials development (Huang, 2024), and the demographic background information of our participants was similar except for gender.

To investigate the third question, we invited five English teachers at middle schools to a semi-structured interview to investigate their acceptance of the application of LLMs in developing reading materials based on the reading textbooks. The background information of the five English teachers can be found in Appendix D.

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

Before our interview, we also sent the reading materials developed by AI to English teachers. Also, we showed English teachers how to develop their reading materials based on their needs. As such, they can understand the process and effectiveness of AI-assisted materials development.

Moreover, the technology acceptance model (TAM for short) (Venkatesh et al., 2003) was used to guide the semi-structured interviews, which comprised five key elements, including “perceived usefulness”, “perceived ease of use”, “attitude”, “behavioral intention”, and “actual use” (Venkatesh et al., 2012). The questions during the semi-structured interview were based on these perspectives (see Appendix D).

The reason for using this model was that it is dedicated to the user acceptance of technology, which can help us better understand the feasibility of AI-assisted materials development for language teaching. We focused on the first four elements since they are the key perspectives in understanding human perception of AI-assisted materials development.

The interview was conducted online via the Tencent meeting. The English teachers were invited to speak in Chinese, which is their first language. Therefore, they can express their perceptions freely. The duration of each teacher was about 30 minutes. Then, we transcribed and analyzed the recording using the thematic analysis method in NVivo 12 based on the guidelines suggested by Nowell et al. (2017) to enhance trustworthiness and transparency. Guided by the third research question and TAM theory, the first author carefully read through all the transcripts and then generated 20 initial themes. The second author coded the transcripts again to ensure the essential themes were coded. At the outset of the coding procedure, the

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

inter-coder agreement reached approximately 0.85 (85%). Then, they negotiated with each other to synthesize the 20 initial themes and reach a consensus on the eight themes to be reported as visualized in Figure 4.

To enhance the clarity of this paper, we also summarized the relationship between the backward design framework and the research questions in Figure 2, which also showed how each research question was operationalized and investigated by corresponding methods.

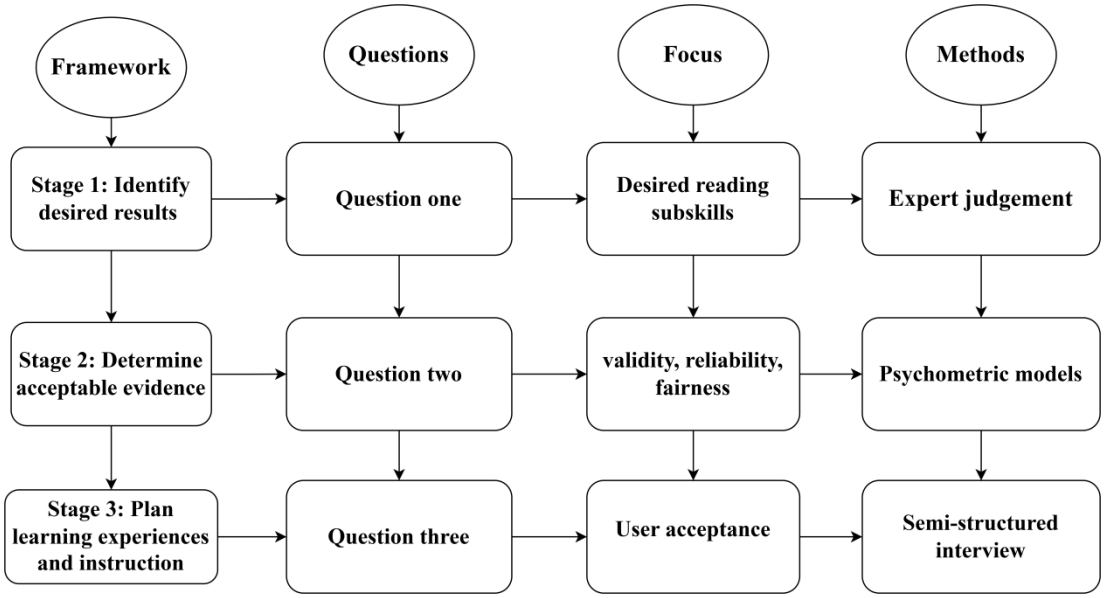

Figure 2. Research roadmap of this study

4. Results

4.1 Result of Desired Reading Subskills

Table 2. Results of expert judgment

Items	Reading subskills	Items	Reading subskills
item 1	A1	item 11	A1
item 2	A2	item 12	A2
item 3	A3	item 13	A3
item 4	A4	item 14	A4
item 5	A5	item 15	A5
item 6	A1	item 16	A1
item 7	A2	item 17	A2

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

item 8	A3	item 18	A3
item 9	A4	item 19	A4
item 10	A5	item 20	A5

Note: A1 to A5 stand for the five reading subskills reviewed in section 2.2, respectively.

Table 2 presents the results of expert judgment based on the operational definitions of the five reading subskills shown in Table 1. The initial Fleiss' Kappa values for these five subskills were 0.65, 0.72, 0.70, 0.73, and 0.72, respectively. Items with lower agreement were then further discussed and refined to reach consensus. The results indicate that the five applied linguistics experts agreed that the reading materials generated by AI could measure the targeted reading subskills as specified in the prompt, based on the operational definitions provided in Table 1.

4.2 Result of Acceptable Evidence

4.2.1 Item Difficulty

Table 3 is the result of item difficulty of each item and the value of standard error (SE) of parameter estimation. The overall item difficulty of AI-generated reading materials ranges from -3.197 to 0.612. It can be seen that most of the SE were around 0.12 to 0.15, which suggests that the estimate of the difficulty parameter was credible. According to the interpretation guideline of item difficulty suggested by Baker and Kim (2017), the difficulty item should be retained within -3 to 3. Generally, most of these items were within the acceptable range, except for item 2.

Table 3. Item difficulty of each item

Item	Difficulty	SE	Item	Difficulty	SE
item1	0.119	0.122	item11	-1.563	0.155
item2	-3.197	0.286	item12	-1.662	0.159

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

item3	0.254	0.123	item13	-0.286	0.123
item4	-1.156	0.139	item14	-0.549	0.126
item5	-1.425	0.149	item15	-1.425	0.149
item6	-1.215	0.141	item16	0.059	0.122
item7	-1.276	0.143	item17	-2.484	0.212
item8	0.239	0.123	item18	-0.728	0.129
item9	-0.045	0.122	item19	0.612	0.127
item10	-0.440	0.125	item20	-0.711	0.129

Note: SE stands for standard error

One fundamental consideration in materials development is that the difficulty of the materials should be adaptive to student ability. The item-person (see Figure 3) in the item response theory can illuminate this issue. As depicted in Figure 3, the reading ability of our participants was normally distributed (mean = 0, SD = 0.42). More importantly, the difficulty of the reading materials was parallel to the reading ability of our participants as a whole. This finding suggests that the item difficulty is adaptive to students' reading ability as a whole, including both difficult and easy items for them to answer.

Concerning reliability, it was found that the marginal reliability is 0.76. According to the interpretation guideline suggested by de Ayala (2022), this can be interpreted as acceptable reliability, supporting the acceptability of the materials developed by AI from the perspective of reliability.

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

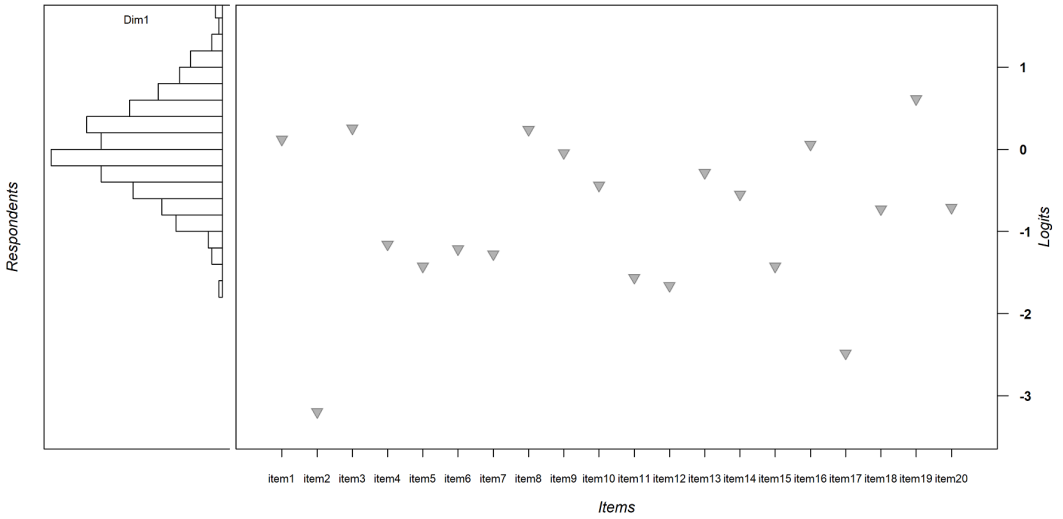

Figure 3. Item-person map

4.2.2 Fairness Analysis

Table 4. Differential item functioning

Items	Stat.	P-value	Adjusted p-value
item1	0.02	0.88	1.00
item2	0.21	0.65	1.00
item3	0.09	0.76	1.00
item4	0.00	0.95	1.00
item5	0.38	0.54	1.00
item6	0.00	0.95	1.00
item7	0.03	0.86	1.00
item8	0.19	0.67	1.00
item9	1.17	0.28	1.00
item10	1.95	0.16	1.00
item11	1.25	0.26	1.00
item12	1.69	0.19	1.00
item13	0.33	0.56	1.00
item14	0.09	0.77	1.00
item15	1.12	0.29	1.00
item16	0.12	0.73	1.00
item17	1.58	0.21	1.00
item18	1.47	0.23	1.00
item19	0.00	0.97	1.00
item20	3.02	0.08	1.00

Apart from the adaptability of difficulty, potential bias towards different genders is also

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

an essential concern in AI-assisted materials development (Huang, 2024), which pertains to the fairness of AI-generated reading materials. Table 4 shows the result of differential item functioning across different genders based on the Mantel-Haenszel test. The key to interpreting the result is to check whether the adjusted *p*-values were significant. The result indicated that none of the *p*-values of the 20 items was significant at the 0.05 level, suggesting that the reading materials generated by LLMs were fair for different genders.

4.3 Result of Teacher Acceptance

In addition to the acceptability of the material generated by AI based on psychometric models, teacher acceptance of using AI in materials development also matters. The key themes discovered during the interview were visualized in Figure 4, which showed how English teachers perceived and experienced the use of AI for materials development.

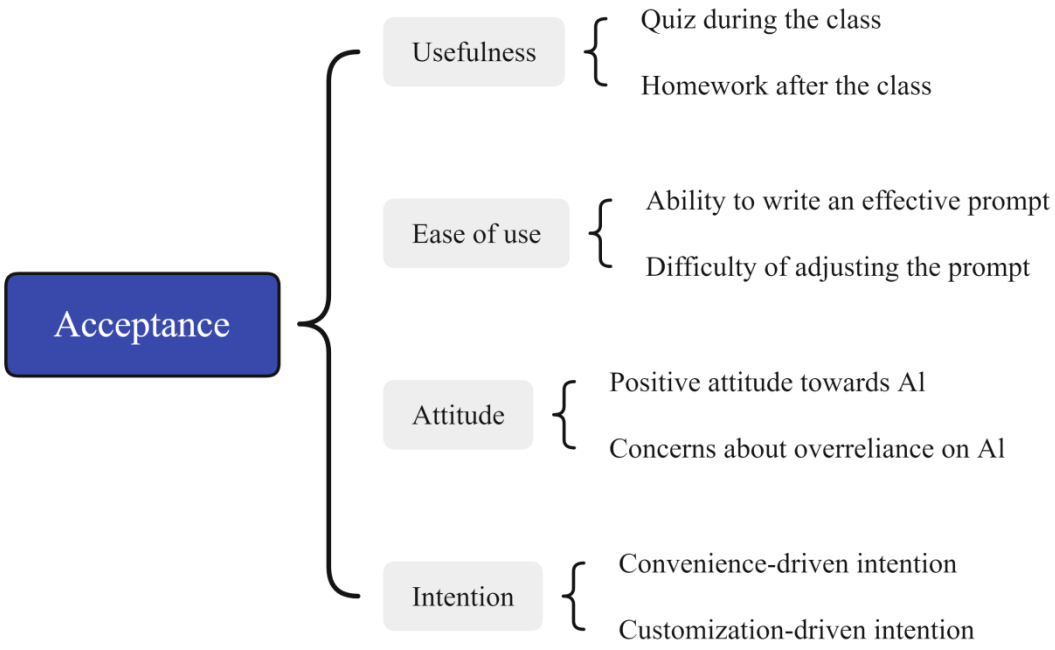

Figure 4. The key themes discovered during the interview

4.3.1 Theme 1: Perceived Usefulness of LLMs

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

English teachers during the semi-structured interviews highlighted their appreciation of the items generated by LLMs. English teachers appreciated the usefulness of AI-generated reading materials for language teaching and learning purposes since these materials are personalized based on the reading materials from textbooks. They felt that the reading materials generated by AI could be used as the materials and quizzes during the class and as the assignment (homework) after the class. Regarding this, teacher 1 noted,

> *"I think the reading materials generated by AI are quite special and useful. The uniqueness of the materials resides in the fact that the materials are personalized and can be customized based on the reading material passage from the textbook. Therefore, these materials can serve as teaching materials during the class and homework for students." (Teacher 1)*

4.3.2 Theme 2: Perceived Ease of Interacting with LLMs

The perceived ease of interacting with LLMs relates to teacher AI literacy in interacting with LLMs. Teachers perceived that the ease dwells in their ability to write an effective prompt and the ability to adjust the prompt for desired outputs. Hence, the readiness to adopt AI tools in teaching practices hinges significantly on teachers' confidence in applying generative AI. An experienced teacher highlighted this point, stating:

> *"The idea of integrating AI into our teaching is exciting, but the actual process of using these tools can be somewhat challenging for those of us who are not versed in digital technologies. It is not just about the willingness to use AI but also about the ability to do so." (Teacher 2)*

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

4.3.3 Theme 3: Attitudes Towards AI-assisted Materials Development

Most of the teachers during the interview embraced the use of LLMs in materials development. There is a general acknowledgment among educators of the transformative potential of AI in personalizing learning experiences. However, one teacher expressed concerns about overreliance on AI. For example, Teacher 3 said,

> *"Personally speaking, I like the use of AI for reading materials development since it can help English teachers create customized teaching materials. However, I wonder if teachers would overly rely on AI technologies in the long term." (Teacher 3)*

4.3.4 Theme 4: Behavioral Intention to Use AI for Materials Development

The behavioral intention included efficiency-driven intention and customization-driven intention. The teachers invited to this study expressed the view that they would like to use LLMs for customizing their own reading materials during the process of course preparation. The reason is that LLMs can increase the efficiency of developing ELT materials. Besides, teachers' workload can be greatly reduced through the application of generative AI. This breakthrough addresses a critical challenge that has long plagued educators—the demanding and often unfeasible task of designing homework that caters to the diverse learning profiles within a classroom. For example, teachers 4 and 5 said,

> *"The application of LLMs in materials development greatly saves me a lot of time and energy in developing reading exercises for course preparation. I used to spend hours designing exercises for my students, but now I just need to give specific prompts to LLMs, and I can customize the teaching materials efficiently." (Teachers 4 and 5)*

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

5. Discussion

5.1 Discussion of Desired Reading Subskills

The first question investigates whether GPT-4o can develop the reading materials for the desired reading subskills in the teaching objectives, which corresponds to the first stage of the backward design framework.

The result indicated that GPT-4o can assess the desired reading subskills as specified in the prompt during the process of materials development. This result is inspiring, which suggests that AI models, such as GPT-4o, can be used to generate reading materials based on our personal needs. This finding was consistent with the findings of Lin & Chen (2024), who explored the capability of LLMs to automatically generate construct-driven reading items for students at the B1 level.

From the perspective of the teaching objectives, these reading subskills were classified into lower-order and higher-order subskills, ranging from “understanding the word meaning” and “understanding explicitly stated information” to “evaluating the author’s attitude” and “summarizing the main idea”. These different levels of reading subskills could represent different levels of teaching objectives according to Bloom’s taxonomy (Anderson, 2005; Krathwohl, 2002). The findings of this study indicated that LLMs have the potential to generate reading items serving language teaching objectives. Compared with the items from textbooks, using these construct-driven items can better evaluate teaching effectiveness.

Therefore, this study has significant pedagogical implications for language teaching. Although this study piloted the reading materials by treating them as a reading test, the utility

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

of the materials developed by AI is not limited to language assessment. For instance, the items for different reading subskills (e.g., summarize the main idea) can be used as quizzes during the class, which can help English teachers evaluate students' comprehension of the text.

Apart from the five reading subskills evaluated in this study, some other reading subskills can also be taken into consideration. For instance, "distinguishing between facts and opinions" and "drawing conclusions" can also be conducted. Investigating other reading subskills can examine the generalizability of this study and lend support to the feasibility and flexibility of LLMs in developing customized reading materials.

Additionally, the reading materials developed by AI can also serve as a diagnostic reading assessment to provide diagnostic feedback to students. The reason is that these reading materials developed by AI were designed to measure different reading subskills on purpose. Therefore, it is possible to provide diagnostic feedback for students based on AI-generated reading materials, which have important pedagogical value for language learning and teaching. For the sake of demonstration, this study also generated a diagnostic reading report based on the AI-generated items (see Appendix E). The diagnostic feedback based on the AI-generated reading materials has the potential to inform students of their strengths and weaknesses in the reading subskills. Future studies can investigate the extent to which the diagnostic feedback is helpful to students.

Furthermore, as the number of items of the reading materials generated by AI is limited, we also replicate this study by generating another 20 items based on the reading passage from the textbooks with GPT-4 (see Appendix G). Although these 20 items have not been piloted,

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

we feel that the quality of the items is very similar to the items tested in this study, which could enhance the generalizability and credibility of the findings in this study.

5.2 Discussion of the Acceptability of Materials

The second question investigated the acceptability of materials generated by AI, which maps to the second stage of the backward design framework. The result showed that the reading materials developed by AI were acceptable in terms of item difficulty. This finding was encouraging, which suggested that teachers can use generative AI to develop reading materials at appropriate difficulty levels for students.

The possible reasons why the item difficulty was adaptive to students were as follows. The first reason might be that we set the requirement in the prompt for LLMs, requesting that the item difficulty should be A2 level. The second possible reason was that the materials used for item generation were adopted from students reading textbooks. The item difficulty in reading comprehension closely relates to the readability of the reading materials. This finding also has important pedagogical implications. This finding also implies that teachers can use the methods used in this study to develop reading materials for students based on their needs.

Besides, LLMs can also be used to adapt the reading materials. For example, teachers can fine-tune the readability of the reading material with LLMs. With this method, teachers can better manipulate the item difficulty (adaptability) of items for students. This can increase the flexibility of the method for developing reading materials for students in different contexts and different targeted language learners.

Additionally, it is also worth discussing that the reading materials were not adaptive to

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

every single student since some students have higher reading abilities, while others might not. However, we feel that the difficulty of the reading materials was adaptive at the group level. One possible solution to further enhance the adaptability of reading materials generated by AI is to develop an adaptive learning system of reading (Kabudi et al., 2021).

Nevertheless, the result also indicated that items 2, 12, and 17 were relatively easy for students' reading ability. One possible reason might be that these items all evaluated the subskill "understanding the explicitly stated information", which belongs to the lower-order subskill (Ilc & Stopar, 2014). Therefore, these items were relatively easier compared with the higher-order subskills, such as "making inferences" and "summarizing the main idea". This finding can also be explained by the greater degree of overlap between the stimuli and the options in the item targeting "understanding explicitly stated information", which has been identified as a key factor influencing listening task difficulty (Cai et al., 2025). This result also contributes to the theoretical understanding of AI-based materials development. Specifically, items targeting higher-level reading subskills, as specified in LLM prompts, are likely to be more difficult than those assessing lower-level reading subskills.

Apart from the acceptability of difficulty, whether the item is biased toward certain groups of students is also of great importance in the backward design framework. If the reading materials were biased, it would be inappropriate to use them in language learning and teaching. The result of differential item functioning suggested that the reading materials generated by LLMs are unbiased towards different genders. This is a positive finding, suggesting that LLMs did not suffer from the potential drawback of gender bias in this study.

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

However, potential biases associated with AI-generated content might also be exhibited in different aspects when we use LLMs to generate reading materials. For example, studies showed that cultural backgrounds and nationality might also be potential biases of LLMs (Chen et al., 2024). This suggests that future studies can further delve into this issue by investigating other variables, such as socioeconomic status, urban/rural, and native language. At present, one of the typical methods used to combat potential biases is reinforcement learning. From the perspective of language teachers, one practical way is to use human review during the process of materials development with LLMs.

5.3 Discussion of Teacher Acceptance

The third research question aims to investigate teacher acceptance of AI-assisted materials development from the perspective of the TAM model, which resonates with the third stage of the backward design framework. Although there are several studies on the acceptance of LLMs (e.g., Zheng et al., 2024), this study differs from previous studies in that this study investigated teacher acceptance in the AI-assisted materials development context and investigated whether they know how to interact with LLMs for materials development purposes.

Concerning usefulness, the findings during the interview revealed that teachers believed that the AI-generated reading materials could serve as quizzes during class and homework after class. These are two important application scenarios from the perspective of students. These materials can also be used as reading materials for formative assessment, which can track students' mastery of content during the class over a whole semester. These potential uses supported the usefulness of AI-generated reading materials for language learning and

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

teaching.

Regarding the ease of interacting with LLMs, the result found that the ease is related to teachers' ability to write an effective prompt and adjust the prompt for the desired outcome. This study can help address this challenge and provide effective and actionable prompts designed for materials development for language teachers, which can help English teachers better utilize AI tools in materials development and improve their user experiences. This study also provided video-based tutorials to lower the threshold of AI-assisted materials development (see the supplementary website).

Concerning attitude, the result of the interview with English teachers found that English teachers embraced the use of generative AI to develop reading materials. They thought the generative AI could help them prepare useful learning resources for students. This was consistent with the findings of Hashem et al. (2023), who also reported that generative AI could serve as a teaching assistant to reduce teachers' workload and prevent burnout. It was also consistent with the finding reported by ElSayary (2024), who reported that LLMs can serve as supporting tools in language teaching and learning. However, this study differs from previous studies in that it explores a concrete way to integrate AI technologies for language teaching purposes.

It is also noteworthy that concerns were raised regarding the overreliance on AI in the development of reading materials in language teaching. This finding highlights the importance of teacher agency in the process of adopting AI for materials development. It also underscores the need to strengthen teachers' agency in evaluating, adapting, and refining AI-generated draft

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

materials (Uyar et al., 2026). Such practices can mitigate potential negative effects and reshape the enactment of teacher agency in AI-assisted reading materials development.

Concerning the intention to use, the interview found that all teachers expressed the intention to use the LLMs for language teaching. Their intentions were mainly driven by convenience and efficiency. This finding implies that LLMs have the potential to address the challenges in materials development, as reviewed and discussed in the literature section. That is, the process of materials development is time-consuming, which is one of the major challenges for language teachers in practice (Yıldız & Harwood, 2023). With the availability of LLMs, the reading textbooks can be further tailored, and English teachers can develop new adaptive learning resources for students using educational technology efficiently.

Moreover, future studies can also investigate other desired tasks, such as cloze items (Yang et al., 2021) and short-answer questions (Huang & He, 2016). Alternatively, English teachers or textbook writers can also use LLMs to generate other useful reading materials based on the textbook for classroom activities. Therefore, it is possible to provide customized reading materials for students in the classroom to enhance student satisfaction and experience with generative AI. As such, we also demonstrated how to generate other types of materials (e.g., matching headings task and summary completion task) based on the reading passages selected in this study, which can enhance the diversity of reading materials for language learning and teaching purposes. The materials can be found in Appendix F. We also provided a tutorial for readers and English teachers to replicate this study, which can be found on the supplementary website.

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

Last but not least, although teachers generally acknowledged the use of AI in materials development, some concerns about the limitations of AI were also worth discussing. Since generative AI is trained based on given data, the content generated by AI might also be hallucinated in practice. Therefore, it is of vital importance for English teachers to be aware of the potential drawbacks of the reading materials developed by generative AI (Yan et al., 2023). This kind of digital literacy is of great significance to ensure that the LLMs are appropriately used in language education. This also prompts our thinking about teacher development in this AI era. English teachers should also update their digital literacy to better collaborate with AI technologies (Jeon & Lee, 2023).

6. Conclusion and Implications

This study investigated the capability of LLMs to develop personalized reading materials within the backward design framework. The study found that GPT-4o could generate customized reading materials for different reading subskills, and the item difficulty of these generated items was adaptive to students' reading ability in general. Besides, the reading materials developed by AI were unbiased towards different genders. Additionally, from the perspective of teachers, they held a positive stance toward the use of LLMs for reading materials development for language teaching. The result of this study lent supporting evidence to the application of LLMs in boosting materials development.

However, several limitations also exist in this study, which merit further investigation. First, this study only examined the potential biases of LLMs using the differential item functioning of genders. Other important factors, such as nationality, cultural background, and

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

discipline, also need to be investigated. Furthermore, this study only investigates the English language. The method applied in this study might also be generalizable to other languages other than English. Future studies replicate this study under different language contexts and develop listening materials with multimodal LLMs (Lin, 2025). In addition, the English teachers invited to this study were homogeneous and limited. Teachers with different backgrounds might have different perceptions, which deserve our further investigation. Moreover, this study has only investigated the reading materials generated by AI. Future studies can also examine the development of audio and audio-visual materials based on multimodal AI.

Despite these limitations, this study has important implications for technology-enhanced materials development. The first implication is that this study shows how the teaching objectives of reading courses can be integrated into generative AI. English teachers can select the reading subskill from the teaching objectives to develop customized learning materials for students. The second implication is that this study provides examples and methods to help English teachers better leverage AI technology to develop supplementary reading materials efficiently. The innovative methods demonstrated in this study can also benefit textbook writers by helping them design more targeted and personalized reading materials economically and productively. The third implication is that this study can help English teachers enhance the usefulness and diversity of the reading materials, which can provide diagnostic reading feedback to students as well.

Note: All the supplementary materials of this study can be found on the OSF platform, which can be downloaded at the link: https://osf.io/vyknx/files/osfstorage

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

References

Anderson, L. W. (2005). Objectives, evaluation, and the improvement of education. *Studies in Educational Evaluation, 31*(2), 102–113. https://doi.org/10.1016/j.stueduc.2005.05.004

Ahmed, A., Jamil, E., Abubakar, M., Batool, A., Akhtar, M. M., Nasiri, M. I., & Ullah, M. (2024). Harnessing the power of ChatGPT to develop effective MCQ-based clinical pharmacy exams. *Journal of Research on Technology in Education*, 1–11. https://doi.org/10.1080/15391523.2024.2425435

Aryadoust, V. (2020). A review of comprehension subskills: A scientometrics perspective. *System, 88*, 102180. https://doi.org/10.1016/j.system.2019.102180

Baker, F. B., & Kim, S. H. (2017). *The basics of item response theory using R*. Springer.

Bamford, J., & Day, R. R. (1998). Teaching reading. *Annual Review of Applied Linguistics*, *18*, 124–141. https://doi.org/10.1017/S0267190500003512

Bouckaert, M. (2015). Designing a materials development course for EFL student teachers: principles and pitfalls. *Innovation in Language Learning and Teaching, 10*(2), 90–105. https://doi.org/10.1080/17501229.2015.1090994

Cai, H., Yan, X., Chuang, P., Pan, Y., & Huo, M. (2025). What makes listening comprehension difficult?: A feature-based machine learning approach to understanding item difficulty. *Applied Linguistics*. https://doi.org/10.1093/applin/amaf079

Chen, X., Aryadoust, V., & Zhang, W. (2024). A systematic review of differential item functioning in second language assessment. *Language Testing*. https://doi.org/10.1177/02655322241290188

Chen, L., (2013). 八年级英语(下册) [English for 8th grade learners, second semester]. Foreign Language Teaching and Research Press.

Chen, X., & Wu, D. (2024). Automatic generation of multimedia teaching materials based on generative AI: Taking Tang poetry as an example. *IEEE Transactions on Learning Technologies*. 1–14. https://doi.org/10.1109/tlt.2024.3378279

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

Curiel, L. C., & Durán, L. G. (2021). A historical inquiry into bilingual reading textbooks: Coloniality and biliteracy at the turn of the 20th century. *Reading Research Quarterly*, *56*(3), 497–518. https://doi.org/https://doi.org/10.1002/rrq.315

de Ayala, R. J. (2022). *The theory and practice of item response theory*. The Guilford Press.

Denny, P., Khosravi, H., Hellas, A., Leinonen, J., & Sarsa, S. (2023). Can we trust AI-generated educational content? Comparative analysis of human and AI-generated learning resources. https://doi.org/10.48550/arXiv.2306.10509

ElSayary, A. (2024). An investigation of teachers' perceptions of using ChatGPT as a supporting tool for teaching and learning in the digital era. *Journal of Computer Assisted Learning, 40*(3), 931–945. https://doi.org/10.1111/jcal.12926

Fernández, C. (2008). Reexamining the role of explicit information in processing instruction. *Studies in Second Language Acquisition*, *30*(3), 277–305. https://doi.org/10.1017/S0272263108080467

Freeman, D. (2014). *Techniques and principles in language teaching* (3rd ed.). Oxford University Press.

Gilmore, A. (2007). Authentic materials and authenticity in foreign language learning. *Language Teaching, 40*(2), 97–118. https://doi.org/10.1017/S0261444807004144

Grabe, W., & Stoller, F. L. (2011). *Teaching and researching: reading*. Routledge.

Hashem, R., Ali, N., Zein, F., Fidalgo, P., & Abu Khurma, O. (2023). AI to the rescue: Exploring the potential of ChatGPT as a teacher ally for workload relief and burnout prevention. *Research and Practice in Technology Enhanced Learning, 19*, 023. https://doi.org/10.58459/rptel.2024.19023

Huang, S. Y. (2024). Gender and diversity in EFL textbook dialogues: Interactional structure and pedagogical implications. *TESOL Quarterly*. https://doi.org/10.1002/tesq.3310

Huang, Y., & He, L. (2016). Automatic generation of short answer questions for reading comprehension assessment. *Natural Language Engineering, 22*(3), 457–489. https://doi.org/10.1017/S1351324915000455

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

Ilc, G., & Stopar, A. (2014). Validating the Slovenian national alignment to CEFR: The case of the B2 reading comprehension examination in English. *Language Testing, 32*(4), 443–462. https://doi.org/10.1177/0265532214562098

Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies, 28*(12), 15873–15892. https://doi.org/10.1007/s10639-023-11834-1

Jin, T., & Lu, X. (2018). A data-driven approach to text adaptation in teaching material preparation: Design, implementation, and teacher professional development. *TESOL Quarterly*, 52(2), 457–467. https://doi.org/10.1002/tesq.434

Kabudi, T., Pappas, I., & Olsen, D. H. (2021). AI-enabled adaptive learning systems: A systematic mapping of the literature. *Computers and Education: Artificial Intelligence*, *2*. https://doi.org/10.1016/j.caeai.2021.100017

Kim, A. (2014). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing*, *32*(2), 227–258. https://doi.org/10.1177/0265532214558457

Konttinen, M. (2022). Towards more learning-centred English-medium education: promoting the combination of backward design and community of practice in teacher training. *Innovation in Language Learning and Teaching*, *16*(4–5), 381–391. https://doi.org/10.1080/17501229.2022.2064469

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice, 41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Lantolf, J. P., & Poehner, M. E. (2010). Dynamic assessment in the classroom: Vygotskian praxis for second language development. *Language Teaching Research, 15*(1), 11–33. https://doi.org/10.1177/1362168810383328

Lee-Ellis, S. (2009). The development and validation of a Korean C-Test using Rasch Analysis. *Language Testing*, *26*(2), 245–274. https://doi.org/10.1177/0265532208101007

Li, J., & Gao, X.S. (2025). Language teachers' developmental trajectories as materials developers. *TESOL Quarterly*, *59*(3), 1719–1749. https://doi.org/10.1002/tesq.3380

Li, H., & Suen, H. K. (2012). Detecting native language group differences at the subskills level of reading: A differential skill functioning approach. *Language Testing*, *30*(2), 273-298. https://doi.org/10.1177/0265532212459031

Lin, M. (2025). Towards Cognitive Intelligence in Financial Document Analysis: A multimodal LLM framework for risk reasoning and due diligence. *Journal of Language*, 1(2), 189–207. https://doi.org/10.64699/25cmhf7434

Lin, Z. (2024). Prompt Engineering for Applied Linguistics: elements, examples, techniques, and strategies. *English Language Teaching*, *17*(9), 14. https://doi.org/10.5539/elt.v17n9p14

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

Lumley, T. (1993). The notion of subskills in reading comprehension tests: An EAP example. *Language Testing*, *10*(3), 211–234. https://doi.org/10.1177/026553229301000302

Merino, P. M. B. S., & Metila, R. A. (2024). Localizing for learning: a designed teacher training program and its developed localized materials for mother tongue education. *Language and Education*, *38*(6), 1080–1097. https://doi.org/10.1080/09500782.2024.2352427

Norton, J., & Buchanan, H. (2022). *The Routledge handbook of materials development for language teaching*. Routledge Abingdon and New York.

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, *16*(1), 1609406917733847. https://doi.org/10.1177/1609406917733847

Rathert, S., & Cabaroğlu, N. (2021). Teachers as slaves or masters to their coursebooks: An in-depth study on two English language teachers' coursebook utilization. *Language Teaching Research*, 13621688211036239. https://doi.org/10.1177/13621688211036239

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

Richards, J. C. (2013). Curriculum approaches in language teaching: forward, central, and backward design. *RELC Journal*, *44*(1), 5–33. https://doi.org/10.1177/0033688212473293

Scott, J. A., & Nagy, W. E. (1997). Understanding the definitions of unfamiliar verbs. *Reading Research Quarterly*, 32(2), 184–200. https://doi.org/10.1598/RRQ.32.2.4

Shi, Y., Ren, P., Wang, J., Han, B., ValizadehAslani, T., Agbavor, F., Zhang, Y., Hu, M., Zhao, L., & Liang, H. (2023). Leveraging GPT-4 for food effect summarization to enhance product-specific guidance development via iterative prompting. *Journal of Biomedical Informatics*, *148*, 104533. https://doi.org/10.1016/j.jbi.2023.104533

Tomlinson, B. (2012). Materials development for language learning and teaching. *Language Teaching, 45*(2), 143–179. https://doi.org/10.1017/S0261444811000528

Tsagari, D., & Sifakis, N. C. (2014). EFL course book evaluation in Greek primary schools: Views from teachers and authors. *System, 45*, 211–226. https://doi.org/10.1016/j.system.2014.04.001

Uiterwijk, H., & Vallen, T. (2005). Linguistic sources of item bias for second generation immigrants in Dutch tests. *Language Testing*, *22*(2), 211–234. https://doi.org/10.1191/0265532205lt301oa

Uyar, A., Karafil, B., & Karakuyu, A. (2026). Artificial intelligence dependency among educators: a scale development and validation study. *Education and Information Technologies*. https://doi.org/10.1007/s10639-025-13862-5

Venkatesh, N., Morris, N., Davis, N., & Davis, N. (2003). User acceptance of information technology: toward a unified view. *MIS Quarterly*, *27*(3), 425–478. https://doi.org/10.2307/30036540

Venkatesh, V., Thong, J. Y., & Xu, X. (2012). Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. *MIS Quarterly*, 157–178. https://doi.org/10.2307/30036540

Vol.2 No.1 2026
DOI: 10.64699/26GXPJ1206

Wang, Q., & Ping, W. (2023). 30 Years of development of English teaching materials: A bibliometric analysis. *International Journal of Applied Linguistics*, *34*(1), 383–408. https://doi.org/10.1111/ijal.12499

Wiggins, G., & McTighe, J. (2005). *Understanding by design* (2nd Edition). Association for Supervision and Curriculum Development.

Xin, J. J. (2024). Investigating EFL teachers' use of generative AI to develop reading materials: A practice and perception study. *Language Teaching Research*. https://doi.org/10.1177/13621688241303321

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2023). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology, 55*(1), 90–112. https://doi.org/10.1111/bjet.13370

Yang., A. C., Chen, I. Y., Flanagan, B., & Ogata, H. (2021). Automatic generation of cloze items for repeated testing to improve reading comprehension *Educational Technology & Society, 24*(3), 147–158. https://www.jstor.org/stable/27032862

Yıldız, A., & Harwood, N. (2023). Why TESOL textbooks are the way they are: The constraints of writing for a global audience. *TESOL Quarterly*, *58*(2), 909–931. https://doi.org/10.1002/tesq.3261

Yılmaz, Ö. K., & Aydın, S. (2025). The impact of the use of Artificial Intelligence–Generated Materials on reading motivation among EFL learners. *Reading Research Quarterly*, *60*(3). https://doi.org/10.1002/rrq.70016

Zheng, Y., Wang, Y., Liu, K. S.-X., & Jiang, M. Y.-C. (2024). Examining the moderating effect of motivation on technology acceptance of generative AI for English as a foreign language learning. *Education and Information Technologies*, *29*(17), 23547–23575.

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J. (2022). Large language models are human-level prompt engineers. https://doi.org/10.48550/arXiv.2211.01910