Journal of Library and Information Studies 23:2 (December 2025) pp.145-183
https://doi.org/10.6182/jlis.202512_23(2).145

Empowering Elementary Learning: Utilizing Large Language Models to Craft Tailored Textbooks with Expert Insight

Da-Chen Lian[1], Mao-Chang Ku[2], Po-Ya Angela Wang[3], Wei-Ling Chen[4], Shu-Kai Hsieh[5]

Abstract

Large language models (LLMs) have in recent years spurred research across various sectors, owing to their remarkable zero-shot or few-shot performance. This capability has become indispensable for individuals seeking to integrate these language models into their workflows effectively. In this paper, based on in-depth linguistic analyses, we explore the application of an LLM, specifically GPT-4, in generating Chinese language textbooks tailored for grade school students. This encompasses the creation of main lesson texts alongside accompanying Chinese character exercises. Experimental results suggest that the LLM-generated textbook lessons are a viable research direction. The initial outcomes demonstrate the ability of LLM to generate texts of satisfactory quality appropriate for a specified grade level. The contributions of this work include pioneering the quantitative analysis of Chinese language textbooks for native speakers in Taiwan and leveraging an LLM to automatically generate textbook content and accompanying Chinese character exercises targeted at native Chinese speakers, which is a novel approach facilitated by the development of prompts tailored to different language learning levels. The study also conducts quantitative and qualitative comparisons between machine-generated lessons and those developed by educational professionals in Taiwan.

Keywords: Large Language Models; Chinese Language Education; Automatic Textbook Generation; Language Learning; In-context Learning

1. Introduction

Large language models (LLMs) have in recent years garnered significant attention following the release of OpenAI's ChatGPT (OpenAI, 2022). ChatGPT, an LLM equipped with billions of parameters and trained on vast amounts of text data, excels at various tasks, making it a versatile assistant for users. Its proficiency in handling both novel and common tasks has spurred research into leveraging LLMs across different sectors. Within the education sector, researchers have explored the potential roles of LLMs in the classroom (Kasneci et al., 2023). Given the generative nature of LLMs, they serve multiple functions, such as a study helper for students or even an assistant for teachers with planning, preparation, and instruction.

In this paper, we aim to leverage an LLM, specifically, OpenAI's GPT-4, in the creation of Chinese language textbook lessons as well as accompanying Chinese character exercises that are aligned with the learning goals of grade students in Taiwan, who are native Chinese speakers. Our research objectives are two-fold:

[1,2,3,4,5] Graduate Institute of Linguistics, National Taiwan University, Taipei, Taiwan
* Corresponding Author: Shu-Kai Hsieh, E-mail: shukaihsieh@ntu.edu.tw

first, to ascertain whether prompt engineering can guide an LLM to generate culturally relevant and pedagogically sound textbook lessons tailored for Taiwanese elementary school students; second, to determine if the same approach can produce effective Chinese character exercises that address the semantic and phonetic aspects of character composition.

Although prior work has examined automatic textbook generation, most existing approaches focus on English and adopt rule-based authoring or script-driven natural language processing (NLP) pipelines. These systems often rely on predefined templates and lack adaptability across different language systems or educational contexts. Furthermore, cognitive models of language acquisition have rarely been integrated into these generation frameworks. Crucially, Chinese poses unique challenges that distinguish it from alphabetic languages. Its logographic writing system lacks explicit word boundaries, its characters are composed of radicals that convey semantic or phonetic cues, and its syntactic structure differs substantially from that of English. These characteristics complicate automatic text generation and require level-sensitive linguistic control, particularly when tailoring material to specific developmental stages.

To achieve these objectives, we leverage GPT-4's extensive context capabilities and include in the prompt descriptive statistics that capture important characteristics of existing textbooks, linguistic knowledge that accounts for the key linguistic features found in textbook lessons, as well as guiding principles of the makeup of textbooks that are found in Taiwan's Curriculum Guidelines of 12-year Basic Education. We utilize prompt engineering to present our information in a more conducive form for the model to generate satisfactory output.

The contributions of this work encompass several pioneering endeavors in the field of Chinese language education. Firstly, this study introduces a groundbreaking quantitative analysis of Chinese language textbooks and utilizes an LLM to automatically generate textbook content and associated Chinese character exercises. This innovative approach is made possible through the development of level-aware prompts tailored to specific proficiency levels, enhancing the adaptability and efficacy of the generated materials. Furthermore, the research conducts comprehensive quantitative and qualitative evaluations, comparing the machine-generated lessons with those meticulously crafted by education professionals in Taiwan. Through these multifaceted contributions, this study significantly advances the understanding and application of LLMs in Chinese language education, opening up new avenues for future research.

2. Literature Review

2.1 Large language models and emergent abilities

Language models, particularly those based on the transformer architecture (Vaswani et al., 2017) with its self-attention mechanism, have revolutionized language processing. The Generative Pre-trained Transformer (GPT) series exemplifies this, starting with GPT demonstrating unsupervised pre-training's potential (Radford et al., 2018). Subsequent iterations like GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020) progressively scaled up model parameters

and datasets, enhancing generalization and leading to GPT-3's robust few-shot learning capabilities without fine-tuning (Brown et al., 2020).

The sheer scale of these models led to the observation of *emergent abilities*, which are not present in smaller models and appear unpredictably once a certain threshold of size or training compute is met, rather than through simple extrapolation of scaling laws (Wei, Tay, et al., 2022). Performance of these abilities can seem random until this threshold, after which it improves significantly. Key emergent abilities relevant to LLM utility include multi-step reasoning, such as chain-of-thought (CoT) prompting (Wei, Wang, et al., 2022), and crucially, instruction following (Wei, Tay, et al., 2022). Instruction following, vital for effective user interaction with LLMs designed as assistants, exhibited emergent behavior: smaller models fine-tuned for it performed worse, while models above a certain size significantly improved.

Recognizing that raw capabilities are not inherently aligned with human values (e.g., truthfulness, helpfulness; Ouyang et al., 2022), researchers developed methods that incorporated user feedback. Ouyang et al. (2022) introduced InstructGPT, models often preferred over GPT-3 due to such alignment training. OpenAI later launched ChatGPT, utilizing similar feedback-driven methodologies to achieve widespread success (OpenAI, 2022). This spurred further development of advanced LLMs like Meta's Llama series (Touvron, Lavril, et al., 2023; Touvron, Martin, et al., 2023), Google's Gemini (Gemini Team Google, 2023), and Anthropic's Claude (Anthropic, 2024), all excelling as user assistants.

2.2 Bias in LLMs

Biases, such as gender bias, can emerge in LLMs from the training data even after filtering efforts (Rae et al., 2022; Raffel et al., 2020; Sheng et al., 2019), and can be subtle (e.g., benevolent bias; Rae et al., 2022; Raffel et al., 2020). For instance, Wan et al. (2023) found ChatGPT and Alpaca generated recommendation letters with gender-stereotypical word choices, more formal/agentive language for males, and hallucinated biases. Detecting and mitigating such biases is crucial for educational materials (Wan et al., 2023), with AI guardrails like Llama Guard (Inan et al., 2023) offering potential solutions.

2.3 LLM tokenization and the Chinese writing system

Modern multilingual LLMs often show strong performance in Chinese, with some Chinese-developed models outperforming others on local tasks (Z. Chen et al., 2025; DeepSeek-AI, 2025; Lee, n.d.; Yang et al., 2025). One of the limitations is tokenization.

Tokenization is needed to convert words into tokens found in a language model's vocabulary, which can then be mapped into a numerical format that the language model can understand (Hugging Face, n.d.). A popular tokenization method is byte pair encoding (BPE) (Radford et al., 2019; Sennrich et al., 2016), which creates a fixed-size subword vocabulary by merging frequent byte pairs, enabling the processing of rare words and avoiding out-of-vocabulary issues (Mielke et al., 2021; Radford et al., 2019). While byte-level BPE can sometimes produce tokens corresponding to Chinese radicals (e.g., GPT-4o's tokenizer shows a 94% correlation for initial tokens that share semantic radicals), its frequency-based and non-

semantic nature can mislead LLMs (Haslett, 2025). Haslett (2025) found LLMs (GPT-4, GPT-4o, Llama 3) sometimes drew incorrect conclusions about character meanings from sub-character tokens, for instance, being more likely to equate shared initial tokens with shared semantic radicals even when it is incorrect.

2.4 LLMs and Taiwanese cultural awareness

Previous efforts in Taiwan have aimed to imbue LLMs with local social and cultural knowledge, exemplified by projects like TAIDE (Llama-based, pretrained on Taiwanese data; https://taide.tw/index), Taiwan LLM (Llama 2-based with continual pretraining/fine-tuning; Y.-T. Lin & Chen, 2023), and the Breeze model family (Mistral/Llama 3-based with expanded Traditional Chinese vocabulary/fine-tuning; Hsu et al., 2024; MediaTek Research, 2025). However, training on Taiwanese data can introduce or reinforce local biases. Hsieh et al. (2024) explored gender and ethnicity biases in 3 Traditional Chinese LLMs (TW-LLM, Breeze, TAIDE), finding small but present gender biases (e.g., TAIDE associates females with lower ability/traditional roles) and more pronounced ethnicity biases (e.g., TW-LLM associates Indigenous People with athletic giftedness). The researchers also found more toxic language was more likely to be linked to severe anti-female bias.

2.5 Prompt engineering

Prompt engineering, crafting directives for LLMs, is key to optimizing performance (Kulkarni & Bansal, 2023; Sanmarchi et al., 2023; White et al., 2023; Zgreabăn & Suresh, 2023). Strategic prompts elicit specific responses (Y. Chen et al., 2023; Fu et al., 2022; Wei, Wang, et al., 2022), with complexity increasing for intricate tasks. Techniques like CoT prompting (Wei, Wang, et al., 2022), step-back prompting (Zheng et al., 2023), and the Socratic method (Chang, 2023) highlight the importance of structured inputs. Santu and Feng (2023) proposed the TELeR taxonomy (Turn, Expression, Role, Level of Details) for designing complex task prompts, emphasizing that detailed specification—clear goals, data, sub-tasks, evaluation criteria, and explanations—improves LLM performance. Given the complexity of generating nuanced educational materials, this study adopted the structured approach of TELeR. Its principles, especially regarding the "Level of Details," "Role" definition, and clear goal setting, were instrumental in formulating our prompts (elaborated in the Methodology section and Table 11) to systematically guide LLMs in producing educational content aligned with curriculum standards.

2.6 LLMs in education

The capabilities of LLMs have led many to explore how they can be utilized in different fields. Kasneci et al. (2023) discuss the benefits and challenges of using LLMs in education, highlighting their use from elementary to university levels. Elementary students can enhance their critical thinking through generated questions, while university students can summarize texts and extract main ideas. Teachers can also leverage LLMs to create personalized teaching materials and lesson plans, including syllabi and discussion questions, and even to facilitate language teaching by outlining relevant vocabulary and grammar.

Denny et al. (2024) introduced the notion of

Generative AI for Education (GAIED), framing both the promise and the open challenges that LLMs bring to learning contexts. Many studies have begun to map this space: (1) agents driven by LLMs for domain tutoring (Hao et al., 2024) and VR companions that enhance immersion (Zhang et al., 2025), (2) personalized learning support (Hayat & Hasan, 2023), and (3) automatic generation of pedagogical artifacts (Blobstein et al., 2023; Choi et al., 2023). Collectively, these studies outline the rapidly expanding research agenda of GAIED that spans content creation, delivery, and education support.

However, the use of LLMs is not without risks. Issues such as potential copyright infringement and inherent biases from training data are significant concerns. Additionally, educators have reported mixed results with LLMs (Tlili et al., 2023), such as inaccuracies in generated summaries and variability in responses to the same queries, highlighting the need for critical engagement with the technology.

Despite these challenges, LLMs introduce innovative teaching methods that require new skills from educators, particularly how to interact with the technology to optimize output quality. This adaptation is crucial for harnessing the full potential of LLMs in educational settings.

2.7 Taiwan Mandarin textbook curriculum guidelines

Taiwan's 12-year Basic Education curriculum (Ministry of Education [MOE], 2018) provides a comprehensive framework aiming for holistic education and the development of students' core competencies, such as critical thinking and cultural awareness. Our research on creating Mandarin textbook materials for elementary students in Taiwan is structured according to this curriculum. The key components that guided our LLM-based generation of textbook lessons and character exercises are summarized in Table 1. This structured approach provided a robust framework for the LLM to create content that not only teaches language skills but also promotes a broad understanding of cultural and societal dynamics.

3. Methodology

3.1 Textbook

In the dynamic field of education, adapting and optimizing teaching materials is crucial to effectively address the diverse learning needs of students. This research introduces a novel methodology that combines in-depth linguistic analysis of textbooks with the capabilities of LLMs to develop a systematic approach to the generation of textbooks for elementary Chinese language learners.

This study systematically analyzes quantitative aspects of textbook compilations to understand how educational materials meet students' learning needs. Utilizing a dataset of 338 text samples from various Mandarin textbooks in Taiwan, we examine the distribution of vocabulary, grammar, and discourse features to guide an LLM in generating high-quality, linguistically accurate materials. By leveraging LLMs and NLP techniques, this research aims to develop an advanced system for compiling textbooks that support comprehensive language development. The generated texts are evaluated both quantitatively and qualitatively against reference textbooks to further refine the content and ensure alignment with educational

Table 1. Overview of Relevant Taiwan Mandarin Textbook Curriculum Guidelines

Curricular aspect	Details for elementary Mandarin	Relevance to this study
Overall goal	Holistic education; development of core competencies like critical thinking, creativity, communication, and cultural awareness.	Provides context for educational material design.
Educational stages	Three stages aligned with elementary grades: Grades 1–2 (Low Level in this study) Grades 3–4 (Mid Level) Grades 5–6 (High Level)	Defines distinct levels for LLM-generated content and exercises.
Learning dimension: Learning performance	Six areas: Listening, Oral Expression, Phonetic Symbols and Their Application, Character Recognition and Writing, Reading, and Composition.	"Character Recognition and Writing" directly informs the design of character exercises.
Progression in character recognition & writing	Stage 1 (Gr 1–2): Mastering components & radicals. Stage 2 (Gr 3–4): Expanding vocabulary using components. Stage 3 (Gr 5–6): Understanding character structure & meaning.	Guides level-specific design requirements for character exercises.
Learning dimension: Learning content	Three themes: 1. Linguistic Discourse (structural attributes like phonetics, syntax) 2. Textual Representation (genres: Narrative, Lyrical, Expository, Argumentative, Practical) 3. Cultural Connotation (Material, Community, Spiritual Culture)	Themes and their components could guide LLM prompt design for textbook content, ensuring linguistic accuracy, appropriate genre, and cultural depth.
Integrated educational topics	Diverse topics incorporated, e.g., Gender Equality, Human Rights, Environmental Education, Technology Education.	Ensures generated content is well-rounded and addresses contemporary societal aspects.

standards. The next section will introduce the dataset in detail and explore the linguistic phenomena observed across different dimensions.

3.1.1 Dataset

Our dataset, sourced from 3 Taiwanese textbook publishers, includes 338 modern Chinese texts, segmented into Low (grades 1–2), Mid (grades 3–4), and High (grades 5–6) educational levels, with each level containing roughly 100 passages (High: 138). Text length increases with grade level, averaging 162, 491, and 927 characters for Low, Mid, and High levels, respectively. Texts underwent automatic word segmentation, part-of-speech tagging (ckip-transformers; Note 1),

Table 2. Summary of Dataset Features by Grade Level

Features (Avg.)	Low (n = 100)	Mid (n = 100)	High (n = 138)
Text length (characters)	162	491	927
Verb tokens per text	26	73	135
Verb types per text	8	23	46
Noun tokens per text	34	114	220
Noun types per text	8	25	48
Connective tokens per text	1	6	14
Connective types per text	0.3	0.6	0.8
BEI construction frequency (per sentence)	0.006	0.022	0.032
BA construction frequency (per sentence)	0.061	0.039	0.047
Common sentence types	Declarative	Declarative, Exclamatory	Declarative, Interrogative
Predominant scenarios	Environment, Daily Life	Entertainment, Travel, Education	Education, Entertainment, Travel

and word sense tagging (Chinese Wordnet; Note 2), followed by linguistic analysis of vocabulary, grammar, and discourse components. Table 2 provides an overview of key linguistic features across levels.

Vocabulary. Vocabulary complexity, including token and type counts for verbs, nouns, and connectives, generally increases with educational level (Table 3). Table 4 shows the distribution of specific types of these vocabularies. The major types of verbs are VH (state intransitive verbs, which resemble adjectives and convey descriptive qualities), VC (action transitive verbs, requiring two arguments without a prepositional object), and VA (action intransitive verbs). Predominant noun types include Na (general nouns, typically modified by quantifiers but not adverbials), Nb (proper nouns), Nc (location nouns), and Nd (temporal nouns). Among connectives, Caa (coordinating conjunctions) and Cbb (subordinating conjunctions) are the most frequent subcategories. The interquartile range (IQR) and standard deviation (SD) for these word classes are presented in Table 5, indicating greater diversity in word usage at higher levels.

Grammar. In our analysis, special attention is given to the BEI (Note 3) and BA (Note 4) constructions, examining their frequency and linguistic features across educational levels.

As illustrated in Table 6, it is evident that the BEI Construction emerges in texts from lower levels, and both the average frequency per sentence and the overall proportion of occurrences increase with each level. We can also see that the verbs commonly associated with the BEI Construction are predominantly VC, with VH,

Table 3. Average Token and Type Counts for Verbs, Nouns and Connectives per Lesson by Grade Level

Level	Verbs		Nouns		Connectives	
	Tokens	Types	Tokens	Types	Tokens	Types
Low	26.0	8.2	33.7	7.8	1.4	0.3
Mid	73.3	23.2	113.7	25.2	5.5	0.6
High	135.4	45.7	219.8	47.7	13.9	0.8

Table 4. Distribution of Specific Verb, Noun, and Connective Type Counts by Grade Level

Grade	Verb types				Noun types					Connective types				
	VA	VC	VH	Total	Na	Nb	Nc	Nd	Total	Caa	Cab	Cba	Cbb	Total
Low	175	259	283	717	591	39	57	57	744	5	1	0	23	29
Mid	472	813	850	2,135	2,072	122	306	172	2,672	13	2	1	58	74
High	772	1,432	1,745	3,949	3,688	306	564	250	4,808	15	3	2	78	98

Table 5. IQR and SD of the Three Word Classes in Each Level

Level	Verb		Noun		Connective	
	IQR	SD	IQR	SD	IQR	SD
Low	38	92	29	157	11	12
Mid	95	278	91	538	23	27
High	171.8	526	191	959	28	36

Table 6. Distribution of BEI Construction and Associated Verb Types by Grade Level

Grade	Avg. BEI per sentence	Proportion of BEI (%)	Verb type counts in BEI construction			
			VC	VH	VE	VB
Low	0.006	3.1	1	0	0	0
Mid	0.022	32.5	14	5	3	0
High	0.032	64.4	25	4	0	2

VE, and VB appearing in conjunction with the BEI Construction only in the middle and high level.

On the other hand, the BA Construction also appears in texts from lower levels. Interestingly, while the overall proportion of occurrences of BA Construction increases with each level, the average frequency per sentence is higher in lower levels as shown in Table 7. Lastly, similar to the BEI Construction, the verbs most commonly associated with the BA Construction are VC, followed by VG, VD, and VB.

Discourse. In terms of discourse structure, we can first observe the usage of punctuation marks. As shown in Table 8, the 5 most frequently used punctuation marks in textbook materials are commas, periods, parentheses, colons, and exclamation marks, while other punctuation marks such as question marks, dashes, and semicolons are less commonly used.

If we use punctuation marks as a criterion for classifying sentence types, we can observe from Table 9 that the number of declarative sentences is the highest, and shows a gradually increasing

Table 7. Distribution of BA Construction and Associated Verb Types by Grade Level

Grade	Avg. BA per sentence	Proportion of BA (%)	Verb type counts in BA construction			
			VC	VG	VD	VB
Low	0.061	16.5	19	5	5	0
Mid	0.039	31.3	37	0	9	4
High	0.047	52.2	51	7	0	8

Table 8. Overall Punctuation Mark Count

Punctuation mark	Count
Comma	10,706
Period	4,612
Parenthesis	3,328
Colon	1,162
Exclamation	1,053
Pause	739
Question	596
Dash	311
Semicolon	194
Etc.	27

Table 9. Distribution of Sentence Type Counts by Grade Level

Grade	Declarative	Interrogative	Exclamatory
Low	513	73	184
Mid	1,617	226	493
High	2,482	297	376

trend that correlates positively with the level. This is primarily because texts in higher grades tend to be longer. It is worth mentioning that although interrogative sentences follow a similar pattern to declarative sentences with the frequency increasing with grade levels, exclamatory sentences have the highest frequency in middle grades and decrease slightly in higher grades. Table 2 succinctly summarizes key linguistic features, including text length, vocabulary richness, grammar construction frequencies, sentence types, and scenario distributions across grade levels.

Moreover, we utilize the Scenario Wordlist provided by National Academy for Educational Research in Taiwan (National Academy for Educational Research, 2021). This comprehensive list comprises 17 distinct semantic tags representing different scenario categories, including *Personal Information*, *Daily Life*, *Occupation*, *Entertainment*, *Travel*, *Social*, *Body and Health*, *Education*, *Shopping*, *Food*, *Public Services*, *Safety*, *Environment*, *Society*, *Culture*, *Emotions and Attitudes*, and *Technology*. Each semantic tag has a list of its semantic-frame words. With the wordlist, we classify every word in the textbook dataset into its corresponding scenario categories. The aggregated and grade-specific distributions of scenario words are presented in Figure 1 and Table 10. Figure 1 displays these distributions as a 100% stacked bar chart, where each bar represents a text category (i.e., Low, Mid, or High grade-level texts; Titles; or merging the 3 levels into Overall as a category of aggregated text). Within each bar, colored segments illustrate the percentage proportion of each specific scenario category. For precise values, Table 10 provides the exact word counts and corresponding percentages for each scenario within these text categories.

From Figure 1 and Table 10, we can see that in the Low level, the scenario words in the

Figure 1. Scenario Word Category Proportions by Levels, Titles, and Overall

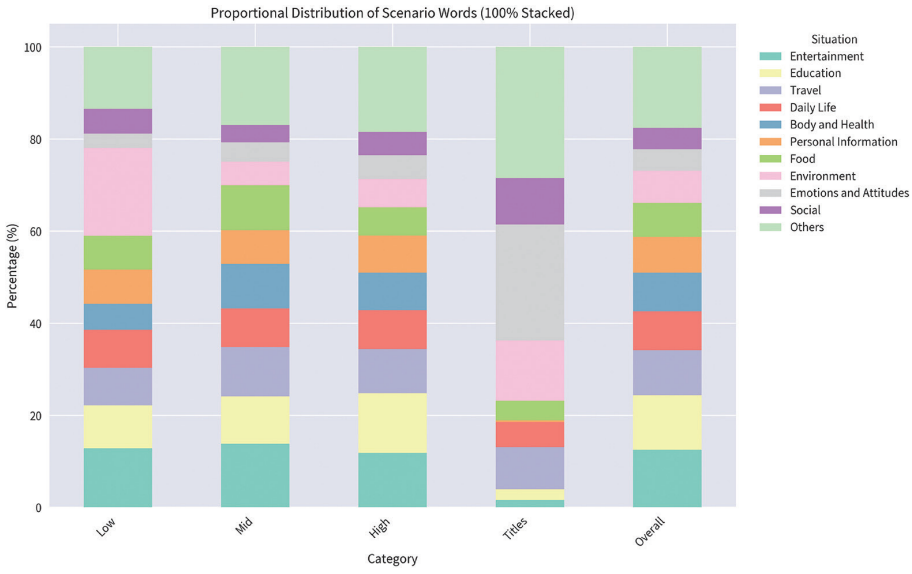


Table 10. Distribution of Scenario Word Counts and Percentages (within each category) by Level, Titles, and Overall

Scenario	Low level (%)	Mid level (%)	High level (%)	Titles (%)	Overall (%)
Entertainment	103 (12.8)	384 (13.8)	632 (11.8)	5 (1.6)	1,119 (12.5)
Education	75 (9.3)	288 (10.4)	696 (13.0)	7 (2.3)	1,059 (11.9)
Travel	66 (8.2)	298 (10.7)	509 (9.5)	28 (9.2)	873 (9.8)
Daily Life	66 (8.2)	234 (8.4)	453 (8.5)	17 (5.6)	753 (8.4)
Body and Health	45 (5.6)	268 (9.6)	435 (8.1)	0 (0.0)	748 (8.4)
Personal Information	60 (7.5)	203 (7.3)	430 (8.0)	1 (0.3)	693 (7.8)
Food	59 (7.3)	272 (9.8)	328 (6.1)	13 (4.2)	659 (7.4)
Environment	153 (19.1)	144 (5.2)	330 (6.2)	40 (13.1)	627 (7.0)
Social	43 (5.4)	104 (3.7)	271 (5.1)	31 (10.1)	418 (4.7)
Emotions and Attitudes	25 (3.1)	116 (4.2)	274 (5.1)	77 (25.2)	415 (4.6)
Occupation	24 (3.0)	87 (3.1)	292 (5.5)	8 (2.6)	403 (4.5)
Safety	20 (2.5)	109 (3.9)	190 (3.6)	0 (0.0)	319 (3.6)
Culture	34 (4.2)	152 (5.5)	127 (2.4)	69 (22.5)	313 (3.5)
Shopping	24 (3.0)	76 (2.7)	195 (3.7)	0 (0.0)	295 (3.3)
Society	3 (0.4)	22 (0.8)	123 (2.3)	4 (1.3)	148 (1.7)
Public Services	3 (0.4)	18 (0.6)	40 (0.7)	0 (0.0)	61 (0.7)
Technology	0 (0.0)	7 (0.3)	17 (0.3)	6 (2.0)	24 (0.3)
Total Words	803 (100.0)	2,782 (100.0)	5,342 (100.0)	306 (100.0)	8,927 (100.0)

texts mainly focus on Environment, which is quite reasonable as it allows younger students to become familiar with their surroundings. As students progress to the Mid level, the distribution of scenario words becomes more even. Entertainment-related scenario words are the most frequent, followed by Travel and Education, with an increasing occurrence of words related to Emotions and Attitudes. In the High level, scenario words related to Education dominate, followed by Entertainment and Travel, and there is a growing presence of words related to Occupations. Additionally, we also categorize the title of each lesson in the dataset based on its contextual relevance to different scenario categories, as shown in the same figure and table. This information is particularly beneficial for our prompt design, which will be discussed in the next subsection.

3.1.2 Prompt design

The process of designing prompts is inherently iterative, necessitating multiple adjustments through trial and error to achieve the intended outcomes. To reduce possible errors, we design

our prompt by leveraging the TELeR framework (Santu & Feng, 2023), focusing on specifying task details, defining context and role, as well as structuring the prompt using the TELeR format (Note 5).

First, we evaluated our prompt with the key factors (task specification details, context, and role) indicated by the framework. Our design is summarized in Table 11.

In our pilot study, the prompt is similar to the narrative braiding case in employing a single-turn instruction style and defining the system role indicated in Santu and Feng's (2023) study. In our updated version, we add more details about how the generated outputs of the model are evaluated since the authors have indicated that this method can improve generation quality. Given that textbook design requires rich knowledge of educational theories, linguistic knowledge, and material writing, the content of the prompt includes additional background knowledge related to the task, which corresponds to the "Level 6" format defined in the authors' framework. The following is our prompt structure designed to elicit the generated (initial) and revised texts (Note 6).

Table A1 (see Appendix A) illustrates how we specify the persona and incorporate background from Curriculum Guidelines and statistics compiled from reference textbooks. Key additional information included:

1. **Inter-level linguistic features of published textbook contents:** Statistics from our textbook dataset analysis were provided to ensure appropriate linguistic elements tailored to the selected level.
2. **The Curriculum Guidelines composed by MOE:** These guidelines informed the structure regarding *Linguistic Discourse*, *Textual Representation*, and *Cultural Connotation*. The level-tailored guidelines for Linguistic Discourse determines word, paragraph, and discourse arrangement of generated texts. Textual Representation contains different styles. We randomly select from Textual Representation and Cultural Connotation for generation.
3. **Writing guidelines for the chosen topic:** A human annotator read the reference textbook contents and assigned each lesson title a *scenario tag* from the Scenario Wordlist (introduced earlier) based on the textual content under that title. These titles subsequently served as the specific topics for the LLM to generate initial and revised texts. Considering that each educational level (Low, Mid, High) exhibits different scenario distributions, each scenario tag corresponds to a distinct list of titles appropriate for that particular level. For content generation, the LLM would first randomly select one scenario tag, and then choose one title from the list associated with that scenario and the target educational level. Each selected scenario then had a corresponding set of writing guidelines (specifying outline and style), which were themselves generated by GPT-4. Diverging styles from these scenario-specific guidelines and the broader Curriculum Guidelines' Textual Representation directives were intended to enhance the diversity of writing within the generated content.
4. **Specified semantic-frame wordlists for the chosen topic:** Level-tailored semantic-frame words were provided to enhance content coherence, which was an improvement over random word selection in the pilot study.

Table 11. Prompt Design Evaluation

Key factors	Definition	Design
Clear goal(s)	Guide the understanding of the LLM and increase the chances of desired output.	Specified target level (Low/Mid/High) with corresponding linguistic features for desired material characteristics.
Associated data	Specify whether the prompt includes data or relies on pre-trained knowledge of the LLM.	Pilot: Provided relevant linguistic knowledge; instructed domain-specific knowledge application. Current: Added Curriculum Guidelines, instructed leverage of this knowledge.
Distinct sub-tasks	Mention multiple steps/sub-tasks clearly to facilitate task division.	Enumerated specific requirements for the LLM.
Evaluation criteria/few-shot examples	Include examples or describe what constitutes a good or bad response.	Pilot: No few-shot examples; constrained responses (POS types, BEI, word limit). Current: Provided one Mid-level text as a one-shot example to guide for high-quality, diverse outputs.
Additional information via information retrieval-based techniques	Use retrieval-based techniques to enhance responses with up-to-date data.	Not strict IR, but provided: 1) Linguistic features from textbook data analysis. 2) Curriculum Guidelines for pedagogical alignment with MOE goals.
Explanation/ justification seeking	Explicitly requesting explanations to understand the reasoning of the LLM for its output.	LLM asked for: explanation/lesson plan per lesson; self-evaluation. Updated version uses TELeR for refined instructions specifying output format/ items more formally.
Defining context and role	Enable the LLM to produce more accurate responses through provided background context to understand a scenario better.	Specified context: role as textbook author, linguistic knowledge of learning levels, textbook editing standards per Curriculum Guidelines.
Expression style	Utilize two primary methods—questions or instructions. The choice between them depends on the requirements of the task and user preference.	Adopted instruction style for directness and efficiency in achieving task goals.
Interaction style	Involve detailed prompt structures, varying between single-turn comprehensive instructions or multi-turn dialogues, significantly affecting LLM performance based on the flow and complexity of the interaction.	Used single-turn direct instructions for achieving task goals.

A one-shot example was provided to address the lack of stylistic variety of the pilot study, and to instruct the model to vary writing styles based on different perspectives and guidelines. Detailed bullet points emphasized tasks and required output formats is shown in Appendix A, Table A2. For revisions, the LLM received the initial text and the same prompt elements (see Appendix A, Tables A1, A2), and was asked for revised text, reasoning, target level, and scores for both versions (see Appendix A, Table A3).

3.1.3 Model

We use OpenAI's GPT-4-0125-preview because of its strong zero-shot performance across tasks and languages with default sampling parameters for all experiments.

3.1.4 Human evaluation

Text set. The study comprised 168 pairs of passages (original + LLM-revised) covering three proficiency targets—Low, Mid, and High.

Presentation. For each pair, the original and the revised version were displayed side-by-side.

Evaluator. One native Chinese-speaking linguist (with a Chinese teaching certificate) served as the evaluator performing the annotation task shown in Table 12.

Dimension-specific cues. When assigning a score, the evaluator also analyzed the following aspects:

Level suitability & readability: Sentence length, idiomatic density, and controlled-vocabulary lists.

Details, conciseness, and tone: Presence/absence of concrete examples, dialogues, and didactic statements.

Coherence & organization: Logical flow, redundant sentences, abrupt topic changes.

Factual & cultural accuracy: Privacy issues, geographical references, and cultural practices.

Language quality: Grammar, idiomaticity, contractions, and quadra-syllabic idioms.

Engagement & tone: Narrative voice and appropriateness of moral or instructional voice.

3.2 Character exercise

3.2.1 Dataset

To develop exercises that support character recognition and construction, we first compiled a list of common Chinese radicals based on entries from the Mandarin Chinese Mini Dictionary (https://dict.mini.moe.edu.tw). To provide structural detail, we used the Chinese Character Gene Dictionary by Bong-Foo Chu (http://open-lit.com/book.php?bid=643), a resource that contains decomposition information for each character specifying a "head" and a "body." These data align with teaching practices found in widely used textbooks (Y.-J. Chen, 2023; C.-P. Lin, 2022),

Table 12. Annotation Tasks and Expected Outputs

Task	Output	Description
T1–Holistic score	1–5 Likert (integer)	Rate the *individual* grades for the revised texts based on whether they meet the level set by the LLM in the recommended-grade-after-revision.
T2–Comparative preference	Categorical	Decide which version is better (*Original, Similar, Revised*). Similar means the two holistic scores differ by ≤ 0.5.

where disassembling and reassembling characters is used as a strategy to build understanding.

For each lesson, we identified the radicals present and selected up to 10 relevant characters for exercise generation. These characters were chosen from the lesson itself or supplemented from the same grade level to ensure sufficient variety. If we found fewer than 10 suitable characters, we included additional ones randomly drawn from the same grade level. All selected characters were then decomposed using the head–body structure to support the next stage of LLM-based exercise design.

Table 13 is an example of how these data are structured. The example represents structured information regarding the radical 門 'door.' It includes:

1. Three textbook characters from the selected lesson that contain the radical 門 'door,' each accompanied by its "head" and "body" components.
2. Extra characters featuring the same radical, drawn from the same grade-level textbook but not originally included in the lesson.

This structured representation supports radical-based instruction by highlighting internal character composition and facilitating their use in targeted learning exercises.

3.2.2 Prompt design

The LLM was prompted to take on the role of a Mandarin Chinese textbook editor, tasked with creating exercises to help Taiwanese elementary school students learn Traditional Chinese characters.

Following the Curriculum Guidelines, the prompt instructed the model to focus on character recognition through the use of common radicals and structural principles. To guide the output format, a sample exercise with answers from an actual textbook was provided. An example is shown in Appendix B.

The exercise unfolds in three parts: character construction, word formation, and sentence construction.

Character construction. In this section, students are given the shared radical 心 'heart' and instructed to combine it with other components—青 'green,' 你 'you,' 田 'field,' and 自 'self'—to form new characters. Each resulting character carries a specific meaning, often related to emotions, social relationships, or mental states, reflecting the semantic contribution of the radical 心 'heart.'

Word formation. After forming the new characters, students are tasked with filling in blanks with compound words using the characters they've created, in this case:

Table 13. Analysis of the Radical 門 'door'

Textbook Characters

Character	Head	Body
們	人	門
間	門	日
開	門	幵

Extra Characters

Character	Head	Body
閱	門	兌
闊	門	活
閉	門	才
問	門	口
閃	門	人
閒	門	月

- 心 'heart' + 青 'green' yields 情 'emotion', fitting into 心情 'mood,' reflecting an emotional state.
- 心 'heart' + 你 'you' produces 您 'you,' used in 您好 'hello,' conveying a formal or respectful greeting.
- 心 'heart' + 田 'field' forms 思 'think,' used in 思考 'ponder,' invoking the act of thoughtful consideration.
- 心 'heart' + 自 'self' results in 息 'rest,' fitting into 休息 'rest,' suggesting a period of relaxation or a break.

Sentence construction. After forming new words, students are presented with a sentence containing multiple blanks, each of which needs to be filled with one of the compound words they previously constructed. The target sentence translates to: "Taking a proper 'rest' during lunchtime not only allows one to relax their 'mood', but also helps one focus more on 'pondering' in class." Students must select the appropriate compound words in a way that makes sense logically and grammatically. This challenges students to consider their meanings not just individually but in a broader context.

For each grade level, one lesson from the LLM-generated textbook was randomly selected for exercise creation. The model received the list of Chinese characters from the lesson, along with radical frequency statistics and head–body structural information for each character. Exercises were then designed around frequently used radicals, selected based on their prominence in the lesson.

The model was further instructed, if necessary, to select additional characters that shared radicals with those in the lesson, drawing from the same grade level. This approach encourages pattern recognition, helping students identify structural relationships among characters based on shared components.

3.2.3 Model

We use OpenAI's GPT-4-0125-preview with default sampling parameters for the character exercise experiments.

4. Results

4.1 Textbook

The LLM generated a total of 52, 60, and 56 texts for the Low, Mid, and High levels, respectively. Analyses of vocabulary, grammar, and discourse were conducted on the initial and self-revised outputs of the LLM, mirroring the methods used for the reference textbooks detailed earlier.

Furthermore, we calculated the Pearson correlation coefficients to assess the alignment of linguistic patterns between the outputs of the model (LLM-generated and LLM-revised) and the reference text corpus across the three levels (Low, Mid, and High), as can be seen in Table 14.

For each linguistic feature (e.g., average verb count), this involved two sets of data points, specifically, the average value of the feature at each level from the reference text corpus and the corresponding average value derived from the LLM-generated (or LLM-revised) corpus. The correlation was then computed between these two sets of three paired values. A high positive correlation suggests that the trend of the linguistic feature across the grade levels in the output of the LLM is similar to that observed in the reference materials. However, it is important to keep in mind that this correlation based on only three data points, while suggesting a trend, has low

Table 14. Pearson Correlation Coefficients for Linguistic Features

Category	Feature	Metric	Reference vs. LLM-generated	Reference vs. LLM-revised
Vocabulary	Verb	Token	0.995	1
		Type	0.952	0.978
	Noun	Token	0.987	0.994
		Type	0.96	0.984
	Connective	Token	0.967	0.984
		Type	0.943	1
Grammar	BEI	Average	0.989	1
		Proportion	0.985	0.995
	BA	Average	0.86	0.364
		Proportion	0.977	0.337
Discourse	Punctuation	—	0.732	0.712
	Sentence type	Declarative	0.994	1
		Exclamation	-0.927	-0.949
		Question	0.125	0.098
	Scenario word	—	-0.001	-0.075

statistical power and cannot be used to alone draw any strong conclusions. This may indicate that the LLM is receptive to descriptive statistics, which we can use to steer the LLM into generating suitable teaching materials (at least from a quantitative perspective).

Analysis of the LLM-generated texts revealed the following results. First, the model exhibited strong performance in both vocabulary and syntax, achieving scores comparable to the reference textbook. Scores further improved after self-revision, indicating the capacity of the LLM for self-refinement. On the other hand, maintaining discourse coherence proved to be more challenging, particularly in handling sentence types beyond simple declarative sentences and in accurately distributing scenario-related vocabulary within the text. Second, the LLM tended to overestimate the complexity level of the generated texts, which was likely influenced by the Mid-level one-shot example provided in the prompt. Lastly, while the LLM-generated scores generally improved after the revision process, human evaluation often favored the original, unrevised texts. This suggests that revisions, while aiming to refine content, may not always align with human judgment of pedagogical effectiveness.

4.2 Character exercise

The LLM-generated character exercises successfully followed prompt instructions designed to reinforce students' comprehension

of Chinese characters through component combination and contextualized use. An example of the generated output is shown in Appendix C.

This exercise demonstrates how the radical-based and head–body structure described in the Methodology is applied in practice. Using 水 'water' as the target radical, the exercise prompts learners to complete partially constructed characters by combining it with components such as 少, 彎, 每, 采, and 满. These combinations yield 沙 'sand,' 灣 'bay,' 海 'ocean,' 深 'deep,' and 滿 'full.'

The resulting characters are then used to construct compound words such as 沙灘'beach,' 海灣 'bay,' and 滿溢 'overflowing,' which appear in a final sentence-completion task. For example, a sentence generated by the model reads as follows:

"As we walked along the ____ in the early morning, we saw a peaceful ____, where the water surface had already ____ due to days of rain."

Learners are expected to insert the newly constructed compound words to complete the narrative, thereby applying vocabulary in a meaningful context and reinforcing lexical recall and expressive use.

While these exercises align with instructional goals and textbook formats, a systematic evaluation of their educational impact remains a direction for future research. We plan to assess both the quantitative outcomes (e.g., completion rates, error types, learning gains across character types) and qualitative feedback (e.g., teacher observations, student reflections) through classroom-based implementation. This will help determine the pedagogical effectiveness of the LLM-generated materials in real instructional contexts.

5. Discussion

5.1 Textbook

Table 15 presents a detailed breakdown of the self-evaluation of the LLM (N = 168 texts), showing how its recommended grade levels shifted after revision, segmented by the original target generation level. Notably, after self-revision, texts tended to converge towards Mid-level (42.3% of all texts were recommended as Mid after revision), which was possibly influenced by the Mid-level one-shot example in the prompt. The LLM perceived significant improvement from its revisions: average self-assigned scores rose from 3.11 to 4.71 (+1.60 points), and 100% of revised texts were rated better than their original versions (Tables 16, 17, 18). Figure 2 contrasts score distributions. In stark contrast, human evaluation showed a slight average score decrease (4.05 to 4.02) and found original texts better or similar in 63.7% of cases (38.1% better, 25.6% similar). This discrepancy suggests LLM revisions, while systemically aligned with its programming, did not consistently match human judgment of quality or suitability for the *revised recommended level* of the LLM (Note 7). Both original and revised texts may have merits for different pedagogical needs.

5.1.1 Qualitative analysis of generated vs. revised texts by level

Low level analysis. At the Low level, the texts presented distinct characteristics before and after revision:

Sentence structure and detail:

- *Originally generated texts* generally featured longer sentences and more detail than expected at this level, which differed from pilot studies where prompts lacked scenario

Table 15. LLM Self-Recommended Grade Level Transitions (Initial vs. Revised) for Texts Grouped by Original Target Generation Level (*N* = 168)

Target level	Initial LLM recommendation	Revised low count (%)	Revised mid count (%)	Revised high count (%)	Total initial count
Low (*n* = 52)	Low (*n* = 52)	42 (80.8)	10 (19.2)	0 (0.0)	52
	Mid (*n* = 0)	0 (0.0)	0 (0.0)	0 (0.0)	0
	High (*n* = 0)	0 (0.0)	0 (0.0)	0 (0.0)	0
Mid (*n* = 60)	Low (*n* = 1)	0 (0.0)	1(100.0)	0 (0.0)	1
	Mid (*n* = 59)	4 (6.8)	49 (83.1)	6 (10.2)	59
	High (*n* = 0)	0 (0.0)	0 (0.0)	0 (0.0)	0
High (*n* = 56)	Low (*n* = 0)	0 (0.0)	0 (0.0)	0 (0.0)	0
	Mid (*n* = 3)	0 (0.0)	1 (33.3)	2 (66.7)	3
	High (*n* = 53)	0 (0.0)	10 (18.9)	43 (81.1)	53
Overall revised distribution (*N* = 168)		46 (27.4%)	71 (42.3)	51 (30.4)	168

Note. This table details how the initial grade level recommendation (Low, Mid, or High) of the LLM for a text changed after self-revision process. These transitions are presented in three main sections, corresponding to the original target level for which the texts were generated. Within each target level section, each cell in the 3x3 sub-matrix provides the count of texts and the number in parentheses is the percentage of all texts that shared that specific initial LLM recommendation within that target group (i.e., row percentage). For example, under Target Low, for texts the LLM initially recommended as Low, the table shows the count and percentage that were subsequently revised to Low, Mid, or High. The final rows of the table summarize the overall distribution of revised recommendations across all 168 texts.

Table 16. Average Overall Scores Before and After LLM Self-revision

Score category	Before revision	After revision	Change
LLM score (Avg.)	3.11	4.71	1.6
Human score (Avg.)	4.05	4.02	-0.03

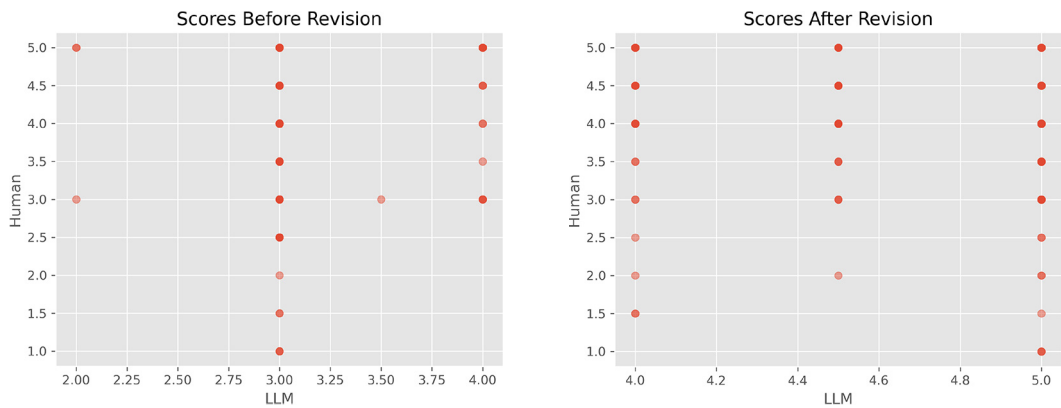
Table 17. Comparative Judgment on Text Quality: Original vs. LLM-Revised Texts

Judgment on revision	Assessed by LLM (%)	Assessed by human (%)
Original text is better	0.0	38.1
Original & revised are similar	0.0	25.6
Revised text is better	100.0	36.3

Table 18. Average Score Changes Within Each Judgment Category for Original vs. Revised Texts

Judgment on revision	Evaluator	*N* (Texts)	Avg. score before rev.	Avg. score after rev.	Avg. score change (Δ)
Original text is better	LLM	0	0	0	0
	Human	64	4.59	3.52	-1.08
Original & revised are similar	LLM	0	0	0	0
	Human	43	4.1	4.1	0
Revised text is better	LLM	168	3.11	4.71	1.6
	Human	61	3.45	4.48	1.03

Figure 2. Comparison of Human and LLM Evaluation Scores



(a) *Before LLM Self-Revision* (b) *After LLM Self-Revision*

guidelines or examples. The increase in detail sometimes made the texts less suitable for beginners.

- *Revised texts* attempted simplification through several methods. Some revisions added more details but resulted in the LLM recommending a higher grade level (Mid) for the text.

Conciseness and language use:

- *Revised texts* often employed Chinese quadra-syllabic idiomatic expressions to achieve brevity suitable for Low-level learners. However, this sometimes increased reading difficulty, making texts more closely resemble classical literature. Another brevity tactic was using contractions, which occasionally led to ambiguity or meaning loss (e.g., "air-water" for "air and water").

Overall suitability:

- Originally generated texts for Low level often had complex sentence structures and extensive detail, potentially exceeding the

needs of beginners. Revisions, while aiming to simplify, sometimes oversimplified to the point of ambiguity or used language constructs that were still too advanced.

Mid level analysis. For Mid-level texts, the comparison between generated and revised versions revealed several key aspects:

Detail, conciseness, and tone:

- *Originally generated texts* were often praised for their detailed examples and conversation-enriched tones, generally meeting Mid or High-level complexity. This richness likely benefited from the scenario-based writing guidelines and the one-shot example provided in the prompt.
- *Revised texts* mostly tended to simplify the text by removing some examples to achieve conciseness or by replacing dialogues with descriptive passages. While revised versions were appreciated for brevity and focus, this sometimes occurred at the expense of valuable details present in the originals. Oversimplification risked losing meaningful content for the target audience.

Real-world knowledge suitability and fact verification:

- The incorporation of real-world knowledge is necessary for content development, but requires careful attention to suitability and factual accuracy. For instance, one of the generated texts described the protagonist's uncle, a doctor, showing a patient's medical records to illustrate his duties; this is a violation of privacy and is inappropriate for educational materials.
- Both originally generated and revised texts sometimes struggled to accurately represent real-world geographical knowledge or cultural descriptions, such as Japanese cultural traditions. These instances suggest a need for further fact verification and more careful consideration of whether fictional elements are beneficial to the educational objectives of the text. While both versions sometimes provided factual supporting details based on real-world entities, multiple verifications are advisable.

Coherence:

- The attempt to incorporate scenario-based writing guidelines specified in the prompts occasionally resulted in an unnatural content flow in both versions.
- Generated and revised texts sometimes concluded with a didactic tone or explicit teaching plan statements (Note 8) specifying moral lessons or concepts, which can feel out of place.
- Nonetheless, the scenario-based guidelines usually helped provide rich supporting ideas for the content.

Language and vocabulary control:

- *Revised texts* were commended for improved control over the vocabulary of different parts-of-speech (POS), indicating a more sophisticated manipulation of language that could benefit learners by exposing them to a proper range of linguistic structures (Note 9). This improvement suggests that the revision step effectively responded to the detailed criteria specified in the prompt.

High level analysis. For texts targeting High-level learners, the following observations were made:

Suitability for target learners and complexity:

- *Originally generated texts* were often more appropriate for advanced learners with their detailed descriptions and longer lengths. They demonstrated the ability of the LLM to generate nuanced and complex narratives.
- *Revised texts* at this level sometimes fell short by omitting details crucial for depth and complexity. These revised versions were often tailored to meet Mid-level needs by simplifying language and reducing details, which could weaken the connection to the theme or topic and make them less suitable for advanced learners.

Detail, conciseness, and tone:

- Sometimes, both versions attempted to mention relevant themes or geographical locations more specifically to strengthen thematic connection and appropriateness.
- *Originally generated texts* were often rich in details, including interactive discussions, making the content more engaging.
- *Revised texts* sometimes stripped away elements that add narrative charm in the pursuit of conciseness. Some revisions included inappropriate word choices or misuses of language. The content could also be affected by scenario-based writing guidelines that led to an excessive focus on certain details, which may deviate from the main theme (Note 10).

Real-world knowledge suitability and fact verification:

- Descriptions involving technology, culture, or geography required verification to remove inaccuracies. However, when sampled, generated texts often supported topics with appropriate real-world examples properly, such as the inclusion of Vancouver's SkyTrain in a Travel scenario text, accurately reflecting real-world cultural context.

Coherence:

- The use of didactic statements at the end of some original texts was not suitable for teaching materials. Revisions often removed these elements, but could result in overly concise content or affect alignment with the given title.

Our findings (summarized in Table 19) highlight that LLM revisions don't guarantee improvement and can lead to oversimplification or loss of engaging details. The tendency of the LLM to overestimate text complexity suggests that prompts need further tailoring, possibly with level-specific few-shot examples, or simpler instructions for Low-level materials.

Compared to existing corpus-based studies of Mandarin textbooks (cf., Y.-J. Chen, 2023; C.-P. Lin, 2022), our work introduces a broader and more systematically structured analysis spanning all three stages of Taiwan's elementary curriculum, incorporating 338 texts, and using CKIP and CWN standards for consistent and fine-grained tagging of verbs, nouns, connectives, and syntactic constructions. Moreover, we introduce scenario-based classification using the Scenario Wordlist (National Academy for Educational Research, 2021), which allows us to analyze how content aligns with semantic domains such as Environment, Education, or Emotions across grade levels. This multidimensional profiling not only enables prompt conditioning for LLM-based generation but also establishes a novel linguistic benchmark that bridges corpus linguistics and AI-driven educational content design. To our knowledge, no prior work in Chinese textbook

Table 19. Issues and Key Challenges across Grade Levels (Condensed)

Grade level	Issues in originally generated texts	Issues after revision	Key challenges
Low	Overly long sentences, excessive detail; unsuited for beginners.	Ambiguous oversimplification; challenging idiomatic expressions for beginners.	Balance simplicity with clarity; avoid ambiguity.
Middle	Rich, engaging dialogues; sometimes overly detailed for optimal readability.	Excessive simplification; removal of valuable dialogues/examples affecting engagement.	Maintain narrative appeal alongside readability and conciseness.
High	Detailed, complex narratives suiting advanced learners; occasional didactic conclusions.	Reduced depth/complexity due to simplification; some inappropriate language use.	Preserve vocabulary complexity; ensure factual accuracy for advanced learners.

research has combined these analytical layers to support large-scale, automated content generation.

For future improvements, prompts should explicitly provide distinct, level-specific examples to guide the LLM accurately. For instance, simplified and direct instructions without examples may be optimal for Low-level texts, whereas carefully chosen and detailed examples might help retain the desired complexity and depth for Middle and High levels. This nuanced approach will help balance the simplicity, narrative appeal, and complexity required at each educational stage.

5.2 Character exercise

This experiment highlights the potential of GPT-4 to generate structured and pedagogically informed exercises for Chinese character instruction. The prompt design focuses on the combination of radicals and components, reflecting morphological principles established in early character learning. However, several limitations emerged during the generation process.

First, guiding the model to generate characters with specific internal structures—particularly those involving head–body decomposition—remains difficult. Even with explicit formatting and examples, the output occasionally fails to follow the intended structural logic. This suggests that character-level morphological control during generation remains limited, likely reflecting both the inherent complexity of sub-character structural tasks (cf., Tseng et al., 2018) and the limited sensitivity of the model to such internal configurations. Tokenization likely contributes to this limitation, as discussed in LLM Tokenization and the Chinese Writing System section.

Second, the phonetic dimension of Chinese characters, which plays a crucial role in pronunciation and literacy development, is not yet addressed in the current design. Incorporating phonetic-semantic relationships into future prompt engineering may provide a more balanced and comprehensive learning experience.

While the current implementation demonstrates the feasibility of generating character-based exercises, further refinement is required to ensure semantic coherence, phonetic integration, and overall instructional effectiveness.

5.2.1 Expert evaluation

To provide a preliminary empirical validation of the LLM-generated character exercises, we elicited evaluation from experts by conducting a questionnaire (Note 11). We received responses from 8 certified elementary school teachers in Taiwan. The participants represent a highly experienced cohort: 4 have been teaching for 13 years or more, 1 for 8–12 years, 2 for 4–7 years, and 1 for 1–3 years. The group has broad expertise across all grade levels, with 5 currently teaching High grades (5–6), 2 teaching Mid grades (3–4), and 1 teaching Low grades (1–2). Notably, the participants were evenly split, with 4 having prior experience of using AI tools in their language classes and 4 having no prior experience.

Teachers were asked to rate the AI-generated exercises on five pedagogical dimensions for each grade level (Low, Mid, and High) using a 5-point Likert scale. For each level, a textbook passage sample and accompanying character exercise were presented. A summary of the mean scores for all questions is presented in Table 20.

The comprehensive results underscore a positive reception while also highlighting key areas for customization. Across all grade levels, the exercises received high ratings for using practical, contextual examples ($M \geq 4.38$) and for their effectiveness in helping students understand character structure ($M \geq 4.13$). This supports the core pedagogical value of the generated content.

Furthermore, teachers expressed a strong willingness to incorporate these exercises into their regular teaching ($M \geq 4.25$). The scores for "Textbook Correspondence" and "Suitability for Stage" reveal the importance of adaptation. Both metrics scored lower in Mid and High compared to Low grades, with "Textbook Correspondence" dropping to a mean of 3.75 in the two higher levels. This suggests that while the fundamental approach is sound, the specific content of a

Table 20. Mean Scores from Teacher Survey ($N = 8$) on the Pedagogical Usability of AI-generated Exercises, Rated on a 5-point Scale

Survey question	Low grade (M)	Middle grade (M)	High grade (M)
Exercises correspond well with textbook content.	4.13	3.75	3.75
The character assembly method is suitable for the students' learning stage.	4.25	4.13	4.13
Words and example sentences are practical and have contextual meaning.	4.63	4.38	4.5
This type of exercise helps students understand components and character structure.	4.5	4.13	4.25
I am willing to incorporate this type of exercise into my regular teaching.	4.5	4.25	4.38

Note. Teachers evaluated the suitability of material for each elementary grade level. *M* refers to mean.

generic example is perceived differently across grade levels, reinforcing the central thesis of this paper on the necessity of an LLM-based workflow capable of generating specifically tailored materials. The teachers also provided constructive suggestions for future iterations, including:

Content refinement: Making component choices "more refined" and ensuring example sentences are more relevant to students' daily lives.

Design and interactivity: Improving the visual layout to be more "lively" and, most frequently, adding grids for students to practice writing the character.

These preliminary findings confirm the pedagogical soundness of our approach and provide a clear and data-driven path for refinement, underscoring the need for adaptable, LLM-powered content generation in language education.

6. Conclusion and Limitations

This study presents an initial exploration of using GPT-4 to automatically generate Chinese textbooks for Taiwanese elementary school students. The generation process was guided by the Curriculum Guidelines of 12-year Basic Education in Taiwan, ensuring alignment with official pedagogical standards and developmental appropriateness for young learners at each grade level. The findings demonstrate the feasibility of using GPT-4 to produce structured and level-appropriate learning materials; however, several limitations highlight the need for more critical assessment and further refinement.

First, while GPT-4 was capable of producing relevant and reasonably structured content, its proprietary nature limits transparency and interpretability. The lack of access to its internal decision-making mechanisms makes it difficult to ensure consistent alignment with pedagogical goals. This issue arises not only in radical-based generation, which requires fine-grained control over character structure and semantic composition, but also at the discourse level, where generated passages may lack coherence or deviate from a clear instructional focus. Future work could explore open-source models (e.g., Llama, Mistral, or Qwen) as more transparent and customizable alternatives to improve structural fidelity, semantic precision, and overall organization.

Second, as LLMs continue to evolve, future research can adopt diverse strategies to improve generation quality. For example, the rapidly growing field of prompt engineering offers methods such as ReAct prompting (Yao, Zhao, et al., 2023), which combines reasoning and action by allowing models to interact with external tools during generation. Similarly, techniques such as CoT (Wei, Wang, et al., 2022) and tree-of-thought (ToT; Yao, Yu, et al., 2023) use step-by-step reasoning to guide the model in a more interpretable manner. In addition, agentic frameworks such as AutoGen (Wu et al., 2023) offer a promising direction by supporting multi-step workflows. By separating tasks into stages such as text generation, exercise design, and output evaluation, this modular approach enables task-specific optimization, thereby improving both adaptability and discourse-level coherence. Together, these advances lay the foundation for generating educational content that meets pedagogical needs.

Third, although the lessons and character exercises are modeled on textbook conventions,

their appropriateness for child learners, particularly in terms of age suitability and instructional effectiveness, has not yet been empirically evaluated. Future research should involve collaboration with textbook editors, practicing teachers, and students to validate and refine the quality of LLM-generated materials. In addition, ethical concerns, such as cultural inaccuracies, biased content, or age-inappropriate examples, require more systematic attention. Strategies addressing these issues may include implementing human-in-the-loop review processes and developing or integrating AI guardrails, such as Llama Guard (Inan et al., 2023), to screen for unsuitable materials before classroom use.

In conclusion, this research underscores the potential of LLM-assisted approaches to generating learning materials for Chinese language education. The inherent complexity of the language, combined with the need for pedagogically sound content, calls for continued exploration and refinement. This study paves the way for future research aimed at developing more effective and accessible tools that support teaching and learning in this context.

Data Availability

The datasets generated and/or analyzed during the current study are not publicly available due to copyright issues but are available from the corresponding author upon reasonable request. We thank Pin-er Chen for her assistance with materials for an earlier version of this paper.

Notes

Note 1 This transformers model is an NLP pipeline tool for Traditional Chinese provided by the CKIP team of Academia Sinica: https://github.com/ckiplab/ckip-transformers.

Note 2 Here we use packages and tools of the Chinese Wordnet (CWN): https://github.com/lopentu/CwnGraph.

Note 3 The BEI Construction (被字句) in Mandarin Chinese expresses a passive voice, where the subject is the recipient of an action. Example: 蘋果被他吃了(The apple was eaten by him.)

Note 4 The BA Construction (把字句) highlights the disposal of an object, indicating the subject causes someone or something to undergo an action. Example: 他把蘋果吃了(He ate the apple.)

Note 5 We experimented with two prompts for automatic textbook generation. In the first pilot study, our prompt was in English with no examples (zero-shot) and without using the TELeR structure defined by Santu and Feng (2023). Furthermore, we randomly selected verbs and nouns from a wordlist for the model to generate texts based on a chosen topic. This leads to incoherent content, shorter text lengths, and inappropriate usage of words because the model attempts to create a story from potentially wildly unrelated words. Thus, in this research, we refine our prompting structure and language, give a one-shot example of what a good lesson looks like, and adopt semantic frame concepts when choosing vocabulary for text generation.

Note 6 How self-reflection and revision can improve generation capabilities of LLMs has been suggested in Shinn et al. (2024) and Cheng et al. (2024). The former indicates how self-reflection facilitates human learning by fixing errors based on previous failures. Similarly, self-reflective feedback offers the LLM specific guidance for improvement, thereby enhancing task performance. The latter suggests that generation improvement can be achieved by directly leveraging the output of model within the unbounded generation space. Hence, textbook generation involves two parts: the original generated texts and the revised texts. We aim to compare the differences between these two generated versions.

Note 7 The human evaluator grades the revised texts based on whether they meet the level set by the LLM in the "recommended-grade-after-revision."

Note 8 These teaching plan statements directly explain the objectives or moral lessons of the generated texts. For example, "Students can learn that..."

Note 9 The statistics that support this argument can be found in 4. Results.

Note 10 We suggest that the scenario definition can be refined to avoid inappropriate generated texts. For example, the scenario related to Personal Information, which requires the inclusion of contact information in its description, misleads the LLM to transform an autobiographic story into a resume-like text.

Note 11 A copy of the questionnaire can be found here (text in Chinese): https://osf.io/wv3g9?view_only=deba2713c6844293a309703af836bd97

References

Anthropic. (2024, March 4). *Introducing the next generation of Claude*. https://www.anthropic.com/news/claude-3-family

Blobstein, A., Izmaylov, D., Yifal, T., Levy, M., & Segal, A. (2023, December 15). *Angel: A new generation tool for learning material based questions and answers* [Conference workshop poster paper]. NeurIPS'23 Workshop: Generative AI for Education (GAIED), New Orleans, LA, United States. https://pse.is/86jlkg

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., 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. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 1877-1901). Neural Information Processing Systems Foundation. https://pse.is/86jd7h

Chang, E. Y. (2023). Prompting large language models with the Socratic method. In *2023, IEEE 13th Annual computing and communication workshop and conference (CCWC)* (pp. 0351-0360). IEEE. https://doi.org/10.1109/CCWC57344.2023.10099179

Chen, Y.-J. (2023). *12 years national education second learning stage Mandarin exercises analysis and research of words and phrases* [Unpublished Master's thesis]. National Taichung University of Education. https://hdl.handle.net/11296/d3ry9w (in Chinese)

Chen, Y., Wen, Z., Fan, G., Chen, Z., Wu, W., Liu, D., Li, Z., Liu, B., & Xiao, Y. (2023). MAPO: Boosting large language model performance with model-adaptive prompt optimization. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 3279-3304). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.215

Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., Gu, L., Wang, X., Li, Q., Ren, Y., Chen, Z., Luo, J., Wang, J., Jiang, T., Wang, B., . . . Wang, W. (2025). *Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling*. arXiv. https://doi.org/10.48550/arXiv.2412.05271

Cheng, X., Luo, D., Chen, X., Liu, L., Zhao, D., & Yan, R. (2024). Lift yourself up: Retrieval-augmented text generation with self-memory. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (Vol. 36, pp. 43780-43799). Neural Information Processing Systems Foundation. https://pse.is/86jeas

Choi, J. H., Garrod, O., Atherton, P., Joyce-Gibbons, A., Mason-Sesay, M., & Björkegren, D. (2023, December 15). *Are LLMs useful in the poorest schools? TheTeacher.AI in Sierra Leone* [Conference workshop poster paper]. NeurIPS'23 Workshop: Generative AI for Education (GAIED), New Orleans, LA, United States. https://gaied.org/neurips2023/files/34/34_paper.pdf

DeepSeek-AI. (2025). *DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning*. arXiv. https://doi.org/10.48550/arXiv.2501.12948

Denny, P., Gulwani, S., Heffernan, N. T., Käser, T., Moore, S., Rafferty, A. N., & Singla, A. (2024). *Generative AI for education (GAIED): Advances, opportunities, and challenges*. arXiv. https://doi.org/10.48550/arXiv.2402.01580

Fu, Y., Peng, H., Sabharwal, A., Clark, P., & Khot, T. (2022). *Complexity-based prompting for multi-step reasoning*. arXiv. https://doi.org/10.48550/arXiv.2210.00720

Gemini Team Google. (2023). *Gemini: A family of highly capable multimodal models*. arXiv. https://doi.org/10.48550/arXiv.2312.11805

Hao, Y., Holmes, J., Waddle, M., Yu, N., Vickers, K., Preston, H., Margolin, D., Löckenhoff, C. E., Vashistha, A., Ghassemi, M., Kalantari, S., & Liu, W. (2024). *Outlining the borders for LLM applications in patient education: Developing an expert-in-the-loop LLM-powered chatbot for prostate cancer patient education*. arXiv. https://doi.org/10.48550/arXiv.2409.19100

Haslett, D. A. (2025). Tokenization changes meaning in large language models: Evidence from Chinese. *Computational Linguistics, 51*(3), 785-814. https://doi.org/10.1162/coli_a_00557

Hayat, A., & Hasan, M. R. (2023, December 15). *Personalization and contextualization of large language models for improving early forecasting of student performance* [Conference workshop poster paper]. NeurIPS'23 Workshop: Generative AI for Education (GAIED), New Orleans, LA, United States. https://gaied.org/neurips2023/files/12/12_paper.pdf

Hsieh, H.-Y., Huang, S.-C., & Tsai, R. T.-H. (2024). TWBias: A benchmark for assessing social bias in traditional Chinese large language models through a Taiwan cultural lens. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 8688-8704). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.507

Hsu, C.-J., Liu, C.-L., Liao, F.-T., Hsu, P.-C., Chen, Y.-C., & Shiu, D.-S. (2024). *Breeze-7b technical report*. arXiv. https://doi.org/10.48550/arXiv.2403.02712

Hugging Face. (n.d.). *Tokenizers - Hugging Face LLM course*. Retrieved May 28, 2025, from https://huggingface.co/learn/llm-course/en/chapter2/4

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., & Khabsa, M. (2023). *Llama Guard: LLM-based input-output safeguard for human-AI conversations*. arXiv. https://doi.org/10.48550/arXiv.2312.06674

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., . . . Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning & Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kulkarni, N. D., & Bansal, S. (2023). Application of generative AI for business analyst role. *Journal of Artificial Intelligence & Cloud Computing, 2*(4), 1-5. http://doi.org/10.47363/JAICC/2023(2)187

Lee, J. [jeinlee1991]. (n.d.). *Chinese-llm-benchmark* [Data Repository]. Github. Retrieved May 27, 2025, from https://pse.is/86jk6t

Lin, C.-P. (2022). *A study on Mandarin textbook and teaching Chinese characters for first graders in an elementary school* [Unpublished Master's thesis]. National Taipei University of Education. https://hdl.handle.net/11296/ybu9m3 (in Chinese)

Lin, Y.-T., & Chen, Y.-N. (2023). *Taiwan LLM: Bridging the linguistic divide with a culturally aligned language model*. arXiv. https://doi.org/10.48550/arXiv.2311.17487

MediaTek Research. (2025). *The Breeze 2 herd of models: Traditional Chinese LLMs based*

on Llama with vision-aware and function-calling capabilities. arXiv. https://doi.org/10.48550/arXiv.2501.13921

Mielke, S. J., Alyafeai, Z., Salesky, E., Raffel, C., Dey, M., Gallé, M., Raja, A., Si, C., Lee, W. Y., Sagot, B., & Tan, S. (2021). *Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP*. arXiv. https://doi.org/10.48550/arXiv.2112.10508

Ministry of Education. (2018). *Curriculum guidelines of 12-year basic education for elementary, junior high schools and general senior high schools: Language arts—Mandarin*. https://pse.is/86jgqj (in Chinese)

National Academy for Educational Research. (2021). *Shuo qing hua jing: Hua yu wen ci yu qing jing fen lei* [Contextualising Chinese words: A classification of Chinese word contexts]. https://pse.is/86jh8q (in Chinese)

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

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (Vol. 35, pp. 27730-27744). Neural Information Processing Systems Foundation. https://pse.is/86jhnf

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. OpenAI. https://pse.is/86jhsq

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. OpenAI. https://pse.is/86jhw5

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., Rutherford, E., Hennigan, T., Menick, J., Cassirer, A., Powell, R., van den Driessche, G., Hendricks, L. A., Rauh, M., Huang, P.-S., . . . Irving, G. (2022). *Scaling language models: Methods, analysis & insights from training gopher*. arXiv. https://doi.org/10.48550/arXiv.2112.11446

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(1), Article 140.

Sanmarchi, F., Bucci, A., Nuzzolese, A. G., Carullo, G., Toscano, F., Nante, N., & Golinelli, D. (2023). A step-by-step researcher's guide to the use of an AI-based transformer in epidemiology: An exploratory analysis of ChatGPT using the STROBE checklist for observational studies. *Journal of Public Health, 32*, 1761-1796. https://doi.org/10.1007/s10389-023-01936-y

Santu, S. K. K., & Feng, D. (2023). TELeR: A general taxonomy of LLM prompts

for benchmarking complex tasks. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 14197-14203). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.946

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th annual meeting of the Association for Computational Linguistics (Volume 1: Long papers)* (pp. 1715-1725). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1162

Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3407-3412). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1339

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2024). Reflexion: Language agents with verbal reinforcement learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (Vol. 36, pp. 8634-8652). Neural Information Processing Systems Foundation. https://pse.is/86jjs7

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments, 10*(1), Article 15. https://doi.org/10.1186/s40561-023-00237-x

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and efficient foundation language models*. arXiv. https://doi.org/10.48550/arXiv.2302.13971

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., . . . Scialom, T. (2023). *Llama 2: Open foundation and fine-tuned chat models*. arXiv. https://doi.org/10.48550/arXiv.2307.09288

Tseng, C.-C., Chen, H.-C., Chang, L.-Y., Hu, J.-F., & Chen, S.-Y. (2018). A corpus-based analysis of radical position and the degree of freedom of permissible positions and examination of the influential radical properties. *Bulletin of Educational Psychology, 49*(3), 487-511. https://doi.org/10.6251/BEP.201803_49(3).0007 (in Chinese)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. von Luxburg,

S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 6000-6010). Neural Information Processing Systems Foundation. https://pse.is/86jjyh

Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K.-W., & Peng, N. (2023). "Kelly is a warm person, Joseph is a role model": Gender biases in LLM-generated reference letters. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 3730-3748). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.243

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). *Emergent abilities of large language models*. arXiv. https://doi.org/10.48550/arXiv.2206.07682

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (Vol. 35, pp. 24824-24837). Neural Information Processing Systems Foundation. https://pse.is/86jjzx

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A prompt pattern catalog to enhance prompt engineering with chatgpt*. arXiv. https://doi.org/10.48550/arXiv.2302.11382

Wu, Z., Chen, J., Liu, J., Shan, Y., Ma, X., Shukla, T., Yang, F., Hsu, H., Qin, Y., Meng, Y., Ratner, A., Dow, S., & Xiong, C. (2023). *AutoGen: Enabling next-gen LLM applications via multi-agent conversation*. arXiv. https://doi.org/10.48550/arXiv.2308.08155

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., . . . Qiu, Z. (2025). *Qwen3 technical report*. arXiv. https://doi.org/10.48550/arXiv.2505.09388

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

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). *ReAct: Synergizing reasoning and acting in language models*. arXiv. https://doi.org/10.48550/arXiv.2210.03629

Zgreabăn, B.-M., & Suresh, R. (2023). Prompting ChatGPT to draw morphological connections for new word comprehension. In M. Hardalov, Z. Kancheva, B. Velichkov, I. Nikolova-Koleva, & M. Slavcheva (Eds.), *Proceedings of the 8th student research workshop associated with the international conference recent advances in natural*

language processing (pp. 98-107). Incoma. https://aclanthology.org/2023.ranlp-stud.11.pdf

Zhang, Z., Zhang-Li, D., Yu, J., Gong, L., Zhou, J., Hao, Z., Jiang, J., Cao, J., Liu, H., Liu, Z., Hou, L., & Li, J. (2025). Simulating classroom education with LLM-empowered agents. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies (Vol. 1: Long papers)* (pp. 10364-10379). https://doi.org/10.18653/v1/2025.naacl-long.520

Zheng, H. S., Mishra, S., Chen, X., Cheng, H.-T., Chi, E. H., Le, Q. V., & Zhou, D. (2023). *Take a step back: Evoking reasoning via abstraction in large language models*. arXiv. https://doi.org/10.48550/arXiv.2310.06117

(Received: 2025/3/18; Accepted: 2025/7/30)

Appendix A
Supplementary Materials

Table A1. Prompt Design Examples for Clarifying Task Goals and Background Knowledge to Generate Teaching Texts

Prompt section	Partial prompt example (English translation)
Description of high-level goal	You are a seasoned elementary school textbook author with a deep understanding of Taiwanese Mandarin and Taiwanese culture. Thus, you are capable of creating materials that meet the specific needs of Taiwanese elementary students and resonate with them ... As an author, you have the ability to simplify complex language concepts and make them accessible to learners at different levels. Your cultural sensitivity allows you to appropriately integrate real-life examples and scenarios related to daily life in Taiwan into the teaching materials, fostering a connection between language learning and real-life situations. ... Below is the language knowledge that will assist you in composing texts for different levels, ... "Low Level" (Elementary School Grades 1–2), "Middle Level" (Elementary School Grades 3–4), and "High Level" (Elementary School Grades 5–6).
Additional relevant information gathered	To design teaching materials for {randomly selected level e.g., "Middle" level students (grades 3–4 in elementary school)}, the following guidelines from the middle level of Chinese language textbooks should be followed: Your task is to design textbook content based on data statistics, the Curriculum Guidelines, various lesson material requirements, and other encouraged incorporated topics. The learning content of the teaching materials is divided into three main themes: "Linguistic Discourse," "Textual Representation," and "Cultural Connotation." Please plan and design the textbook content according to the learning content requirements for the respective grade level. "Linguistic Discourse" includes: - Phonetic symbols: ... - Vocabulary: "1. Recognition of the shape, sound, and meaning of 1,800 common characters... 3. The phonetic and semantic functions of common radicals and components. ... Recognition of 3,000 common terms... 9. The use of measure words... 13. Stories about calligraphy masters." - Sentences and Paragraphs: "1. Usage of various punctuation marks..." - Discourse: "1. Meaningful paragraphs. 2. The main idea, theme, and simple discourse structure. 3. Stories, children's poems, modern prose, etc." "Textual Representation" includes: - Type: "Narrative texts" - Description: "Texts that narrate about people, events, time, places, and things." - Details: "1. The structure of narrative texts. 2. Chronological and flashback narrative techniques." "Cultural Connotations" includes: - Category: "Spiritual culture" - Description: "The artistic, religious, and philosophical connotations contained in various texts." - Details: "The artistic, religious, and philosophical connotations in various texts."

(continued)

Table A1. Prompt Design Examples for Clarifying Task Goals and Background Knowledge to Generate Teaching Texts (contiune)

Prompt section	Partial prompt example (English translation)
Additional relevant information gathered	Requirements for compiling lesson materials: 1. For the low and middle levels, materials can be self-written or edited from existing works; for advanced levels, selections should include important works from both domestic and international classic works, with consideration for continuity into middle school. ... 2. The number of lessons per volume is not strictly defined and should be adjusted according to the depth and length of the selected texts. ... 3. The writing of materials should align with reading comprehension strategies, with the low level emphasizing oral expression and literacy; the middle and high levels emphasizing vocabulary and sentence patterns, and the reading of paragraphs and discourses. ... Additional topics that can be integrated include: {topics listed in the Curriculum Guidelines composed by the MOE} Furthermore, be mindful that the text content may reflect diverse cultural experiences or backgrounds (including different professions, or perspectives of individuals with disabilities). Care should be taken to use different writing styles according to genre and situational requirements.

Table A2. Prompt Design Examples for Providing a One-shot Example, Listing Sub-tasks, Stating Output Formats, and Offering Evaluation Policies to Generate Textbook Lessons

Prompt section	Partial prompt example (English translation)
Example(s)	If needed, you may refer to the writing style of middle level textbook authors but do not directly copy sentences from these texts: /{writing sample}.
A bulleted list of sub-tasks	Please design the teaching material according to the following requirements: 1. The article should be around {selected text length e.g., 268} words. 2. The title of the article is: {selected title e.g. "A Delicious Lesson"} 3. The scenario of the article is: {selected writing guide e.g., "Dining and Cooking" = depicting food, restaurants, and dining experiences, sharing cooking insights. Through narrative or lyrical prose, skillfully blend the food cultures of different countries with personal feelings, narrating smoothly.} 4. The scenario vocabulary of the article includes: {Scenario wordlist e.g., ["bowl", "tea", "fish"].} Please use these words. 5. There should be about {textbook statistic e.g., 114} nouns. 6. There should be about {textbook statistic e.g., 74} verbs. 7. There should be about {textbook statistics e.g., 6} connectives.} 8. The article should include the use of the passive voice marker "BEI."
Output format requirements and ask for explanations on outputs	The output should be presented in JSON format, including the following key values: - design-reasoning: Explain the rationale behind your design of the teaching material. - text: The text is written according to the requirements.
A guideline on how the LLM will be evaluated	You will be graded based on the following criteria: 1. Whether the output matches the target level (low, medium, high). 2. The overall score of the content (1-5, with 5 being the best), based on: - verb: Whether the verbs in the content meet the requirements. - noun: Whether the nouns in the content meet the requirements. - connective: Whether the connectives in the content meet the requirements. - culturally-relevant-content: Whether the content reflects Taiwanese Mandarin, Taiwanese culture, and real-world customs from around the globe. - diverse-content: Whether the content is diverse, with varied topics and writing styles. - meets-requirements: Whether the overall content meets the language knowledge and Mandarin language syllabus requirements.

Table A3. Prompt Design Examples for Stating Output Formats to Evaluate Generated Texts and Generate Revised Texts

Prompt section	Partial prompt example (English translation)
Enumeration of sub-tasks, output format requirements, and ask for explanations on outputs	Here is a text written by a textbook author: {generated text} Does this text meet the requirements of the Curriculum Guidelines and the description of language knowledge at various levels? If not, how would you revise it? Please provide the revised text, an explanation of the revisions, and scores for various aspects of the text after revision (1–5, with five being the best). The output should be presented in JSON format, which includes the following key values: - recommended-grade-before-revision: The recommended grade level (low, medium, high) before revision. - recommended-grade-after-revision: The recommended grade level (low, medium, high) after revision. - revised-text: The revised text. - explanation: Explanations for the revised parts. - overall-bf-score: The overall score before revision (1–5, with 5 being the best). - overall-af-score: The overall score after revision (1–5, with 5 being the best). - scores: A JSON object containing scores (1–5, with 5 being the best) in the following categories, with a score given if the category is applicable, otherwise NA. The scoring categories include: - verb: Whether the content's verbs of the content meet the requirements. - noun: Whether the content's nouns of the content meet the requirements. - connective: Whether the content's connectives of the content meet the requirements. - culturally-relevant-content: Whether the content reflects Taiwanese Mandarin, Taiwanese culture, and real cultural customs around the world. - diverse-content: Whether the content is diverse, with varied themes and writing styles. - meets-requirements: Whether the overall content meets the language knowledge and the Curriculum Guidelines.

Appendix B
Example Exercise from the Textbook

Form a character, and then fill in the appropriate character to make a word.
- 「心」＋「青」＝心__
- 「心」＋「你」＝__好
- 「心」＋「田」＝__考
- 「心」＋「自」＝休__

Finally, complete the sentence with all the newly-created words:
在中午時間適當的「____」，不但能使人放鬆「____」，還有助於課堂上更專注「____」。

Appendix C
Example of the Generated Output

Form a character, and then fill in the appropriate character to make a word:
1.「水」＋「少」=「__」
造詞：「__灘」
造句：昨天我們去了海邊的「__灘」玩耍。
2.「水」＋「彎」=「__」
造詞：「海__」
造句：那個「海__」非常適合潛水。
3.「水」＋「每」=「__」
造詞：「__洋」
造句：「__洋」中生活著各種各樣的生物。
4.「水」＋「采」=「__」
造詞：「__海」
造句：「__海」裡有許多未被發現的秘密。
5.「水」＋「萬」=「__」
造詞：「__溢」
造句：雨後小河裡的水「__溢」了。
Finally, complete the sentence with all the newly-created words:
「當我們在清晨沿著『__』散步時，看到了一個寧靜的『__』，那裡的水面因為連日的降雨已經『__』。」

圖書資訊學刊　第23卷第2期　（2025.12）　頁145-183
https://doi.org/10.6182/jlis.202512_23(2).145

賦權國小國語文學習：利用大型語言模型生成符合專業觀點的客製化教材

Empowering Elementary Learning: Utilizing Large Language Models to Craft Tailored Textbooks with Expert Insight

連大成[1]　古貿昌[2]　王伯雅[3]　陳韋伶[4]　謝舒凱[5]
Da-Chen Lian[1], Mao-Chang Ku[2], Po-Ya Angela Wang[3], Wei-Ling Chen[4], Shu-Kai Hsieh[5]

摘　要

大型語言模型（LLMs）近年憑藉其出色的零樣本（zero-shot）或少樣本（few-shot）學習能力，在各領域的研究中備受矚目，並成為許多專業工作流程中導入AI應用的關鍵技術。立基於深度語言學分析，本研究利用大型語言模型（GPT-4），針對臺灣國小學童自動生成客製化的國語文課文與生字練習。實驗初步結果顯示，模型不僅能產出符合指定年級程度的文本，其品質亦具有高度發展潛力。本研究的主要貢獻在於：首先，我們開創性地對臺灣現行國語文教科書進行量化分析；其次，我們設計了一套適用於不同學習程度的提示詞（prompts），成功利用大型語言模型為中文母語者自動生成教材。研究中亦包含模型生成內容與臺灣教育專家編寫版本之間的量化與質性比較分析。

關鍵字：大型語言模型、國語文教學、教材自動生成、語言學習、上下文學習

[1,2,3,4,5] 國立臺灣大學語言學研究所
Graduate Institute of Linguistics, National Taiwan University, Taipei, Taiwan
*通訊作者Corresponding Author: 謝舒凱Shu-Kai Hsieh, E-mail: shukaihsieh@ntu.edu.tw

註：本中文摘要由作者提供。
以APA格式引用本文：Lian, D.-C., Ku, M.-C., Wang, P.-Y. A., Chen, W.-L., & Hsieh, S.-K. (2025). Empowering elementary learning: Utilizing large language models to craft tailored textbooks with expert insight. *Journal of Library and Information Studies, 23*(2), 145-183. https://doi.org/10.6182/jlis.202512_23(2).145
以Chicago格式引用本文：Lian, Da-Chen, Mao-Chang Ku, Po-Ya Angela Wang, Wei-Ling Chen, and Shu-Kai Hsieh. “Empowering Elementary Learning: Utilizing Large Language Models to Craft Tailored Textbooks with Expert Insight.” *Journal of Library and Information Studies* 23, no. 2 (2025): 145-183. https://doi.org/10.6182/jlis.202512_23(2).145