AI4Reading: Chinese Audiobook Interpretation System Based on Multi-Agent Collaboration

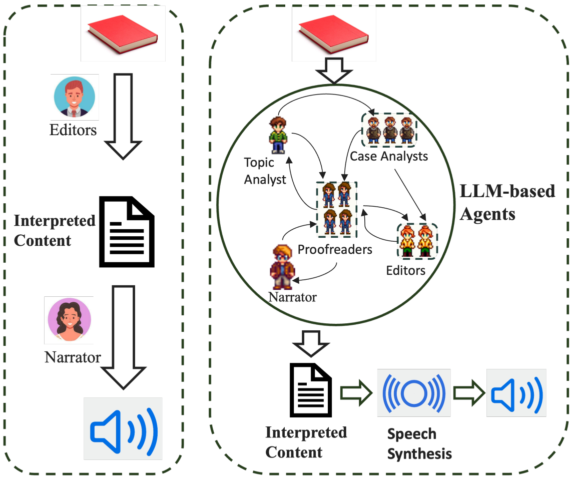
Minjiang Huang[1], Jipeng Qiang[1*], Yi Zhu[1], Chaowei Zhang[1], Xiangyu Zhao[2], Kui Yu[3]
[1]Yangzhou University, [2] City University of Hong Kong, [3] Hefei University of Technology
mz120231035@stu.yzu.edu.cn, {jpqiang, zhuyi, cwzhang}@yzu.edu.cn,
xy.zhao@cityu.edu.hk, yukui@hfut.edu.cn
https://www.ai4reading.top

Abstract

Audiobook interpretations are attracting increasing attention, as they provide accessible and in-depth analyses of books that offer readers practical insights and intellectual inspiration. However, their manual creation process remains time-consuming and resource-intensive. To address this challenge, we propose AI4Reading, a multi-agent collaboration system leveraging large language models (LLMs) and speech synthesis technology to generate podcast-like audiobook interpretations. The system is designed to meet three key objectives: accurate content preservation, enhanced comprehensibility, and a logical narrative structure. To achieve these goals, we develop a framework composed of 11 specialized agents—including topic analysts, case analysts, editors, a narrator, and proofreaders—that work in concert to explore themes, extract real-world cases, refine content organization, and synthesize natural spoken language. By comparing expert interpretations with our system's output, the results show that although AI4Reading still has a gap in speech generation quality, the generated interpretative scripts are simpler and more accurate. The code of AI4Reading is publicly accessible [1], with a demonstration video available [2].

1 Introduction

In recent years, interpretative or retold versions of audiobooks have attracted much attention (Çarkit, 2020; Walsh et al., 2023). Different from unabridged, abridged, or summarized audiobooks, the story is reimagined or modernized to enhance clarity and accessibility for a specific audience, such as younger listeners or those unfamiliar with the original context. This type of audiobook not only preserves the essential themes and narrative arc but also translates archaic language, cultural references, or complex passages into a more relatable and engaging format. To create an interpretative version, publishers and narrators typically collaborate with skilled editors and sometimes the original authors to carefully reword and restructure the text. This process involves identifying and retaining key plot points and character developments while simplifying or rephrasing sections that may be less accessible to modern audiences, which is time-consuming and limits scalability, as shown in Figure 1(a).



(a) Manual Audiobook Interpretation Process

(b) Our Audiobook Interpretation Process

Figure 1: Flowchart of expert-based and LLM-based audiobook interpretation system.

This paper will explore how to use large language models (LLMs) (such as GPT-4o (Achiam et al., 2023) or DeepSeek-V3 (Liu et al., 2024)) to automatically construct an audiobook interpretation system for these categories of books, including psychology, personal growth, business management, etc. By analyzing experts' interpretations, a good audiobook interpretation system should meet the three key objectives:

[*] Corresponding author.
[1]https://github.com/9624219/AI4reading
[2]https://youtu.be/XCLAsRI9v2k

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 211–220
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

(1) Accurate Content Preservation: It must capture and relay core concepts, theories, and strategies in these fields without oversimplification, ensuring the original insights and depth are maintained. **(2) Enhanced Comprehensibility**: The system should transform complex ideas into clear, accessible language, enabling listeners to grasp difficult subjects, and provide more practical cases for explanation. **(3) Logical Narrative Structure**: Maintaining a coherent step-by-step narrative is crucial. This means presenting information in a clear, sequential order that highlights cause-and-effect relationships, so listeners can easily follow the progression of ideas.

Although LLMs have demonstrated strong reasoning capabilities, our tests show that LLMs cannot achieve the above three objectives through chain-of-thought (Wei et al., 2022) or retrieval-augmented generation (Jiang et al., 2023) strategies. Multi-agent systems based on LLMs have gradually risen, showing considerable potential for solving complex problems. They have achieved promising results in fields such as software development (Hong et al., 2023; Nguyen et al., 2024), gaming (Hua et al., 2024; Isaza-Giraldo et al., 2024), and writing (Xi et al., 2025; Bai et al., 2024). Therefore, we will design an audiobook interpretation system based on multi-agent collaboration.

To generate better interpretation manuscripts, we have constructed a combination of 11 agents as shown in Figure 1(b), including: **Topic Analyst** explores book themes, and provides supporting arguments; **Three Case Analysts** expand related knowledge, identify and analyze real-world cases to strengthen the core arguments; **Two Editors** organize content, ensuring logical coherence, clarity, and conversational appropriateness; **Narrator** converts written content into natural spoken language for an improved listening experience; **Four Proofreaders** review and ensure accuracy, logical consistency, and adherence to conversational style.

Finally, our contributions are as follows:

(1) We are the first to study how to automatically construct an audiobook interpretation system using large language models and speech synthesis technology. Compared to manual interpretation, this system, AI4Reading, is not only time- and labor-efficient but also overcomes language barriers, enabling the interpretation of books from different languages into the target language. In terms of system capabilities, our approach provides interpretations of both Chinese books and Chinese interpretations of English books.

(2) For the generation of interpretation manuscripts, we propose a multi-agent collaboration approach. To produce engaging interpretative content, this method considers multiple processes, including topic and case identification, preliminary interpretation, oral rewriting, reconstruction and revision.

(3) We conducted a manual analysis comparing expert interpretations with our results from two aspects: synthesized speech and interpretation manuscripts. The analysis results show that our method produces interpretation manuscripts that are simpler and more accurate. However, the naturalness and appeal of the generated speech are slightly inferior.

2 Related Work

2.1 Audiobook System

The field of audiobook production has evolved to encompass various narration styles, including unabridged, abridged, summarized, and interpretative (or retold) versions.

Traditional audiobooks predominantly focus on unabridged and abridged audiobooks, where unabridged versions deliver the full text as written by the author, and abridged versions condense the narrative to reduce listening time while preserving core themes (Berglund and Dahllöf, 2021). For example, there are tens of thousands of unabridged audiobooks available on Audible [3] and Ximalaya [4] in Chinese. Summarized audiobooks, which distill key ideas and insights into concise formats, have also gained traction, particularly for professional and academic contexts. Blinkist[5] is one of the more popular websites in this category.

More recently, interpretative or retold versions have emerged as a distinct category, wherein the narrative is not merely shortened but is reimagined or modernized to enhance clarity and accessibility for specific audiences (Walsh et al., 2023). This process involves creative editorial adaptations—translating archaic language and complex cultural references into a format that is engaging and relatable, while striving to preserve the original work's essential themes. FanDeng[6] platform in Chinese provides such audiobooks, primarily

[3] https://www.audible.com/

[4] https://www.ximalaya.com/

[5] https://www.blinkist.com/

[6] https://www.fanshu.cn/

narrated by well-known hosts.

Creating interpretative or retold versions of audiobooks is often the most challenging among the formats discussed. It often necessitates collaboration among authors, editors, and narrators to ensure the adapted version maintains the original's essence while resonating with contemporary listeners. In this paper, we will use a multi-agent approach based on LLMs to automatically generate interpretation manuscripts without human involvement.

2.2 Interpretation Generation

Research related to interpretation content generation includes document summarization (Ryu et al., 2024; Ravaut et al., 2024) and document simplification (Fang et al., 2025a,b; Qiang et al., 2023b). In summarization, the goal is to condense content by selecting key points to produce a shorter version of the original text, and in simplification, the objective is to focus on reducing syntactic and lexical complexity to aid readers with varying language proficiencies or cognitive needs (Qiang et al., 2023a,c).

Interpretative generation in this paper requires a creative transformation of the original work: it must reimagine and modernize the narrative to suit a target audience while preserving the essential themes, narrative structure, and nuanced details. This process involves not only removing or rephrasing less essential content but also adding clarifications, restructuring passages, and sometimes even introducing new examples to ensure that the story remains engaging and logically coherent. Such a multifaceted task demands a higher level of domain understanding, creative rewriting, and iterative refinement compared to the relatively straightforward tasks of summarization or simplification.

3 System Design

This section introduces AI4Reading, an intelligent framework for generating interpretive scripts and audio outputs, capable of automatically transforming book content into structured, naturally expressed interpretive scripts and further producing high-quality audio outputs. The system comprises two core modules:

(1) **Interpretation Script Generation:** This module employs a multi-agent collaborative mechanism where specialized roles—such as one Topic Analyst (TA, 🧍), three Case Analysts (from CA1 to CA3, 🧍), four Proofreaders (from PR1 to PR4, 🧍), one Narrator (NR, 🧍), and two Editors (ED1 and ED2, 🧍) —work together to automatically generate the interpretation script.

(2) **Audio Generation:** This module converts the generated interpretive manuscripts into natural, fluent audio outputs by leveraging Text-to-Speech (TTS) technology.

3.1 Interpretation Script Generation

We propose a collaborative multi-agent framework for generating interpretive scripts, as illustrated in Figure 2. This framework takes chapter content as input and leverages specialized system prompts to assign distinct roles and responsibilities to each agent. A detailed description of each stage is presented below.

3.1.1 Topics & Cases Identification (TCI)

This stage mimics human cognitive processes of reading and summarization, distilling core topics and associated supporting cases, which is carried out by three agents: TA, PR-1, and CA-1.

TA processes one chapter S of one book to identify a set of core topics T and a preliminary set of relevant cases C, which is modeled as: $Agent_{TA}(S) \rightarrow (T, C)$, where $T = \{t_1, t_2, \ldots, t_n\}$ is the set of core topics extracted from S, $C = \{c_1, c_2, \ldots, c_n\}$ is the set of preliminary cases associated with the topics, n is the number of extracted topics, with a maximum of 3.

To review whether there are unreasonable topic-case pairs in (T, C), we define an agent Proofreader (PR-1) who rigorously reviews each topic-case pair $(t_i, c_i) \in (T, C)$ in terms of comprehensiveness and relevance. This validation process is defined as: $Agent_{PR-1}(T, C) \rightarrow \{(t, c)\}_{\text{val}} \cup \{(t, c)\}_{\text{inv}}$, where "$\{(t, c)\}_{\text{val}}$" and "$\{(t, c)\}_{\text{inv}}$" denote the valid and invalid topic-case pairs, respectively. If set "$\{(t, c)\}_{\text{inv}}$" is not empty, TA will be called again.

While TA and PR-1 ensure topical relevance and comprehensiveness, preliminary cases often lack depth or context. To fill these informational gaps, we define an agent Case Analyst (CA-1) who enriches the cases with additional background information and key details: $Agent_{CA-1}(S, T, C) \rightarrow (T, C')$. CA-1 ensures that the final output aligns with the original content while effectively supporting subsequent tasks.

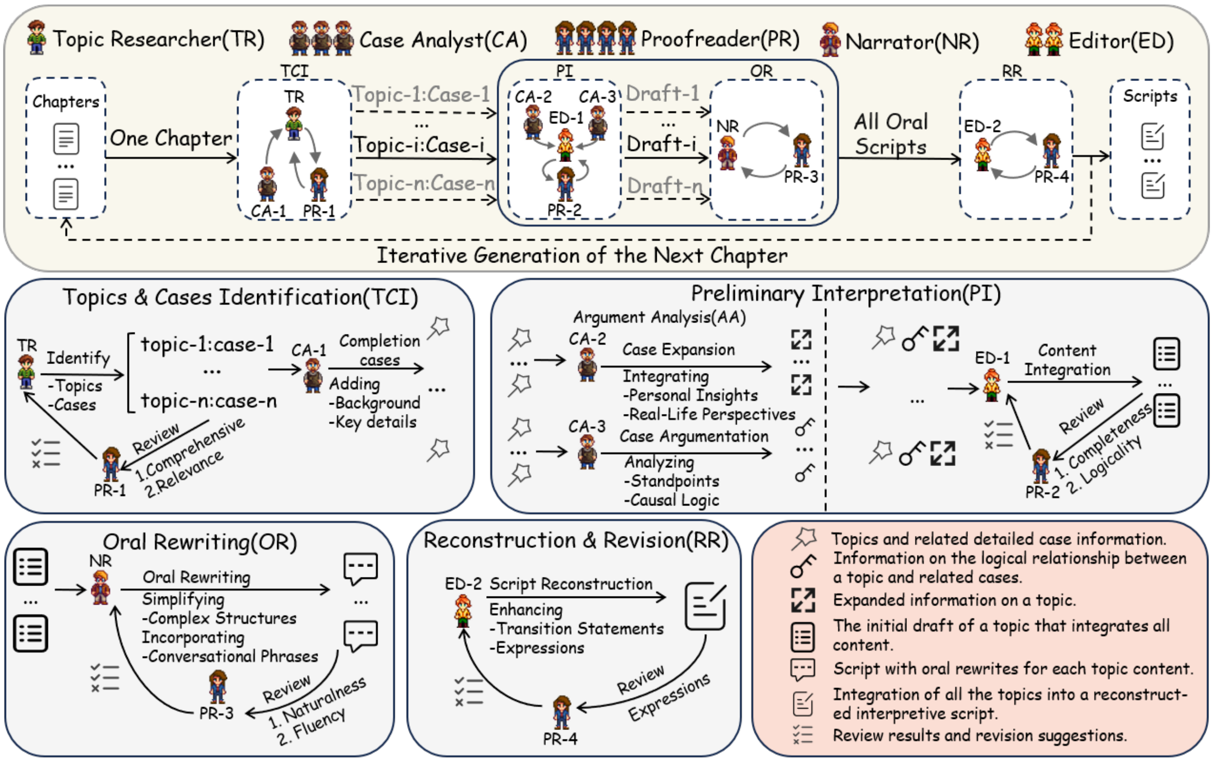


Figure 2: The process of interpretation script generation in AI4Reading based on multi-agent collaboration.

3.1.2 Preliminary Interpretation (PI)

The output (T, C') from the previous stage did not consider how to better facilitate the understanding of theoretical content. In this stage, we aim to supplement each "topic-case" pair by incorporating personal anecdotes and real-life scenarios using these agents (CA-2, CA-3, ED-1, and PR-2), making the content more relatable to everyday life.

CA-2 expands upon each topic-case (t_i, c_i') by integrating personal insights and real-life perspectives: $Agent_{CA-2}(t_i, c_i') \rightarrow a_i$, where a_i represents the expansion for case c_i'. CA-3 constructs logical arguments to demonstrate how each case supports its corresponding topic. This involves analyzing standpoints, causal logic, and ensuring consistency with the chapter content and topics: $Agent_{CA-3}(t_i, c_i') \rightarrow l_i$, where l_i represents the logical argument established for topic t_i.

Considering that (a_i, l_i) lacks the continuity and narrative flow required for a cohesive interpretive manuscript, we define a new Editor, ED-1, who synthesizes all prior analytical findings into a coherent and well-structured preliminary draft for each topic. $Agent_{ED-1}(t_i, c_i', l_i, a_i) \rightarrow d_i$. where d_i is the preliminary draft of topic t_i.

To further ensure rigor and clarity, PR-2 evaluates each topic draft d_i based on two dimensions: completeness and logicality. For drafts that do not meet the required standards, PR-2 provides constructive feedback: $Agent_{PR-2}(d_i) \rightarrow f_i$, where $f_i = (compt_i, log_i, sr_i)$, $compt_i$ and log_i belonging to $\{"Yes", "No"\}$ indicate whether the draft satisfies the completeness and logicality criterion, sr_i contains specific revision suggestions only if $compt_i = "No"$ or $log_i = "No"$, otherwise, $sr_i = \emptyset$.

Based on the feedback f_i, ED-1 iteratively refines the draft d_i to be improved through multiple rounds of optimization:

$$d_i = \begin{cases} d_i & \text{if } sr_i = \emptyset, \\ Agent_{ED-1}(d_i, sr_i) & \text{if } sr_i \neq \emptyset. \end{cases} \qquad (1)$$

The optimization process continues until d_i passes review ($sr_i = \emptyset$) or reaches the maximum number of allowable iterations $I_{\max}$.

3.1.3 Oral Rewriting (OR)

In this stage, two agents, NR and PR-3, refine the draft d_i to make its expression more conversational and easier for the audience to understand.

NR performs a conversational paraphrase of d_i by: (1) simplifying complex structures, and (2) integrating colloquial lexicon and conversational markers, which is modeled as: $Agent_{NR}(d_i) \rightarrow o_i$ where o_i represents the oral script of d_i.

To ensure high-quality oral output, PR-3 conducts rigorous evaluation of oral output o_i, concentrating on two critical dimensions: linguistic naturalness and delivery fluency: $Agent_{PR-3}(\mathcal{O}) \rightarrow G$, where g_i is the feedback of o_i.

Based on the feedback g_i, NR iteratively refines the oral script o_i. This process continues until one of the following termination conditions is met: (1) PR-3 considers the requirements to be met; (2) The maximum number of allowable iterations I_{max} is reached.

3.1.4 Reconstruction and Revision (RR)

Upon completing the oral rewriting of multiple scripts $\mathcal{O} = \{o_1, o_2, \ldots, o_n\}$, the phase moves into the reconstruction and revision phase for the final interpretive manuscript. This stage involves ED-2 and PR-4.

ED-2 integrates all independent interpretive segments $\mathcal{O} = \{o_1, o_2, \ldots, o_n\}$ into a coherent and unified full-length document. The process begins by selecting the first segment o_1 as the initial draft, and subsequent segments $\{o_2, \ldots, o_n\}$ are incrementally incorporated into the current manuscript according to predefined integration guidelines:

$$M_i = \begin{cases} o_1 & i = 1, \\ Agent_{ED-2}(M_{i-1}, o_i) & i > 1. \end{cases} \quad (2)$$

where M_i denotes the manuscript after incorporating the i-th segment o_i. Each integration step ensures logical clarity and natural transitions, continuing until all segments are included: $M = M_n$. Through this process, ED-2 helps the manuscript maintain a seamless narrative flow while preserving the depth and richness of the content.

To prevent inconsistencies during the integration process, we utilized PR-4 to evaluate the entire manuscript M and provide feedback on overall coherence, fluency, and naturalness. The iterative refinement process follows the same mechanism as in the PI and OR stages. Through this series of adjustments, the final interpretation script will achieve better structural coherence and fluency.

3.2 Audio Generation

After the interpretation script is generated, we need a TTS tool to convert the script M into audio. Modern TTS technology not only produces natural and smooth speech but also adjusts the tone and emotional expression according to the content's characteristics, providing listeners with a richer and more vivid auditory experience. In our system, we adopt high-quality Fish-Speech (Liao et al., 2024) as TTS tool.

Additionally, we add transition sound effects at the beginning and end of each chapter to help listeners more clearly perceive the transitions between chapters, thus improving the overall comfort and logical flow of the listening experience. This design not only enhances the user's listening experience but also increases the coherence of the content and the efficiency of knowledge absorption.

3.3 Agent Configuration Rationale

The selection of 11 specialized agents in our AI4Reading framework was a deliberate design choice, stemming from our goal to emulate the collaborative and iterative workflows of human expert teams involved in creating high-quality interpretations. We aimed to decompose the complex task of generating an audiobook interpretation into more manageable, focused sub-tasks, each addressable by an agent with a specific role and set of responsibilities.

Our initial explorations considered simpler architectures, particularly relying on a single, powerful LLM to generate the entire interpretation for a chapter using comprehensive prompts with chain-of-thought (Wei et al., 2022) strategy. However, this approach proved unsuitable for our specific task of interpretation generation for several critical reasons: (1) Tendency towards Summarization, Lacking Interpretative Depth: We observed that even with explicit instructions to "interpret" and "explain," a single LLM often defaulted to producing a high-quality summary of the chapter. This output, while concise and accurate in terms of content distillation, inherently lacked the crucial characteristics of an interpretation. Interpretations require going beyond mere summarization to include elaborations, real-world examples, connections to broader concepts, and a narrative style designed to enhance listener comprehension and engagement, which aligns with our objectives but is contrary to the nature of a summary. (2) Insufficient Content Volume and Coverage: The generated text from a single LLM pass was frequently too brief to adequately cover the nuances and key arguments presented throughout an entire book chapter. Interpretations, by their nature, often expand on the original text to clarify complex points, thus requiring a more substantial word count than a summary. The single-LLM outputs often felt like condensed overviews rather than thorough, engaging explanations.

The current structure, therefore, addresses these shortcomings. Each agent has a clearly defined, relatively narrow scope, allowing for more precise prompting and more reliable execution of its specific function. This granular approach, with iterative feedback loops provided by Proofreader agents, was found to yield more consistent, structured, and truly interpretative scripts that are richer in content, better cover the source material, and more effectively meet the multifaceted requirements of audiobook interpretation. Future work may explore optimizations to this configuration, but the current setup provides a robust foundation to overcome the limitations of simpler, single-pass approaches.

4 Experiments and Evaluation

4.1 Experimental Setup

We will evaluate our system manually from the following two aspects: audio quality and interpretation script.

System Configuration: Our system is built upon MetaGPT (Hong et al., 2023) with Deepseek-V3 (Liu et al., 2024) as the base LLM, with the model temperature set to 1.3, max_token set to 8192, I_{max} set to 3, and the prompting strategy using 0-shot prompting. All used prompts are open-sourced on GitHub.

Benchmark: The competitive product FanDeng[7] (FD) serves as the benchmark for comparison. FD is China's premier knowledge service platform, founded by Dr. Deng Fan, a renowned media scholar, TV host, and communication expert. All the compared contents are narrated by Fan Deng himself.

Data: We selected interpretative books from the FD website as evaluation material, including 10 explanatory excerpts randomly sampled from 10 chapters across five books, covering topics such as personal growth and business finance. The average duration of our system-generated segments was 4 minutes 59 seconds, and the FD-provided segments averaged 4 minutes 33 seconds.

4.2 Evaluation Metrics

To evaluate the quality of the generated speech and interpretation text, we developed an evaluation system [8] where users rate the speech and interpretation text without knowing whether the results are from our method or FD.

We conducted a human evaluation using a 1-5 Likert scale with 7 undergraduate participants (4 male, 3 female) from diverse academic backgrounds (e.g., computer science, engineering, business, etc.). All annotators are Chinese speakers. We recorded the time users spent on each webpage interface in the system backend. Users were unaware of this in advance.

Audio Evaluation: The audio generated by our system and that from FD were presented in a randomized order. Users were asked to listen to the two audio clips sequentially, with the order of presentation also randomized. After listening to each clip, users completed a survey assessing the following three dimensions: (1) Naturalness (Nat.): Evaluates whether the audio sounds natural and fluent. (2) Concentration (Conc.): Assesses whether the user felt fatigued or distracted during the listening process. (3) Comprehension (Compn.): Measures the user's understanding of the audio content.

Interpretation Script Evaluation: Users were initially required to read the original text of the chapters to ensure a thorough understanding of the source content. Users then responded to questions based on the selected script. The evaluation encompassed the following four dimensions: (1) Simplicity (Simp.): Assesses the effectiveness of the interpretation script in reducing the comprehension difficulty of the original text. (2) Completeness (Compt.): Checks whether the interpretation script omitted any key information from the original chapters, as identified by the evaluators. (3) Accuracy (Acc.): Determines whether the main ideas conveyed by the interpretation script were consistent with those of the original text. (4) Coherence (Coh.): Analyzes whether the interpretation script contains any logical inconsistencies, such as abrupt content shifts, broken causal relationships, or contradictions.

4.3 Results

Although there were seven users, we observed that two users had significantly shorter evaluation times, suggesting they may not have completed the tasks conscientiously. Consequently, valid data from only five users were retained for analysis.

[7] https://www.fanshu.cn/

[8] http://49.232.199.229:14444/, username:admin1, password:1admin

Annotators	Methods	Audio Quality			Textual Content			
		Nat.	Conc.	Compn.	Simp.	Compt.	Acc.	Coh.
1	FD	5.0	4.3	2.9	4.2	3.4	4.0	4.2
	OURS	4.2	3.8	3.8	4.2	3.2	4.4	4.8
2	FD	4.7	4.3	4.0	4.0	4.0	4.2	4.0
	OURS	3.9	3.8	3.8	4.6	4.6	4.2	4.8
3	FD	5.0	4.2	3.1	5.0	3.6	3.8	3.8
	OURS	4.6	3.6	2.6	5.0	3.6	4.0	4.0
4	FD	4.8	3.9	2.7	5.0	4.8	4.6	4.6
	OURS	3.6	2.5	1.8	4.8	4.4	4.8	4.2
5	FD	5.0	4.1	4.0	3.6	3.2	4.2	3.8
	OURS	4.2	3.3	3.4	4.6	4.0	4.2	4.2
Average	FD	4.9	4.2	3.3	4.4	3.8	4.2	4.1
	OURS	4.1	3.4	3.1	4.6	4.0	4.3	4.4

Table 1: Results of human evaluation on two dimensions: audio quality and interpretation script. "FD" is from https://www.fanshu.cn/

The results are shown in Table 1. In the audio evaluation, our system achieved better results in terms of simplification, while other metrics were lower than FD's results. However, it is also evident that our system is a highly effective approach for generating speech. Regarding textual generation, our method outperformed in all four metrics, demonstrating that our multi-agent-based approach is highly effective for generating interpretative scripts. The evaluation results fully demonstrate the advantages of multi-agent collaboration in content creation and validate the effectiveness of our framework.

5 Conclusions

In this paper, we propose a novel multi-agent collaborative system for interpretative audiobook generation, addressing the critical challenges of cost, quality, and language accessibility in traditional audiobook production. By simulating the workflow of human-authored interpretation scripts through specialized agents, including Topic Researchers, Case Analysts, Editors, etc. The system achieves efficient and accurate content distillation.

Acknowledgement

This research is partially supported by the National Natural Science Foundation of China (62076217), the National Language Commission of China (ZDI145-71), the Blue Project of Jiangsu province, and the Top-level Talents Support Program of Yangzhou University.

Limitations

Our study acknowledges several limitations that must be addressed in future research. First, the evaluation sample was relatively small, which may not fully capture the diversity of listener experiences. A broader validation involving a larger and more varied group of participants is essential to establish the generalizability and robustness of our system. Second, the current method has been tested exclusively on data from psychology, personal growth, and business management books. This domain constraint limits the system's applicability, as it cannot yet process or generate interpretations for literature or novels. Expanding the system to accommodate these additional genres will be a critical focus of future research efforts.

Ethical Considerations

In developing AI4Reading, we prioritize ethical responsibility in several key areas. First, we ensure that all content generated by the system complies with copyright law and is not used for commercial purposes. Given that our system reinterprets texts, we are committed to maintaining the integrity and core messages of the original works while making them more accessible to listeners. Transparency is another critical factor—we clearly indicate when content is AI-generated to prevent misinformation. Through these measures, we strive to balance innovation with ethical responsibility, fostering trust in AI-driven audiobook interpretation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. Longwriter: Unleashing 10,000+ word generation from long context llms. *arXiv preprint arXiv:2408.07055*.

Karl Berglund and Mats Dahllöf. 2021. Audiobook stylistics: Comparing print and audio in the bestselling segment. *Journal of Cultural Analytics*, 6(3):1–30.

Cafer Çarkit. 2020. Evaluation of audiobook listening experiences of 8th grade students: An action research. *Educational policy analysis and strategic research*, 15(4):146–163.

Dengzhao Fang, Jipeng Qiang, Xiaoye Ouyang, Yi Zhu, Yunhao Yuan, and Yun Li. 2025a. Collaborative document simplification using multi-agent systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 897–912.

Dengzhao Fang, Jipeng Qiang, Yi Zhu, Yunhao Yuan, Wei Li, and Yan Liu. 2025b. Progressive document-level text simplification via large language models. *arXiv preprint arXiv:2501.03857*.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 3(4):6.

Wenyue Hua, Ollie Liu, Lingyao Li, Alfonso Amayuelas, Julie Chen, Lucas Jiang, Mingyu Jin, Lizhou Fan, Fei Sun, William Wang, et al. 2024. Game-theoretic llm: Agent workflow for negotiation games. *arXiv preprint arXiv:2411.05990*.

Andrés Isaza-Giraldo, Paulo Bala, Pedro F Campos, and Lucas Pereira. 2024. Prompt-gaming: A pilot study on llm-evaluating agent in a meaningful energy game. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

Shijia Liao, Yuxuan Wang, Tianyu Li, Yifan Cheng, Ruoyi Zhang, Rongzhi Zhou, and Yijin Xing. 2024. Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis. *arXiv preprint arXiv:2411.01156*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Minh Huynh Nguyen, Thang Phan Chau, Phong X Nguyen, and Nghi DQ Bui. 2024. Agilecoder: Dynamic collaborative agents for software development based on agile methodology. *arXiv preprint arXiv:2406.11912*.

Jipeng Qiang, Yang Li, Chaowei Zhang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2023a. Chinese idiom paraphrasing. *Transactions of the Association for Computational Linguistics*, 11:740–754.

Jipeng Qiang, Kang Liu, Ying Li, Yun Li, Yi Zhu, Yun-Hao Yuan, Xiaocheng Hu, and Xiaoye Ouyang. 2023b. Chinese lexical substitution: Dataset and method. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 29–42, Singapore. Association for Computational Linguistics.

Jipeng Qiang, Kang Liu, Yun Li, Yunhao Yuan, and Yi Zhu. 2023c. ParaLS: Lexical substitution via pretrained paraphraser. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3731–3746, Toronto, Canada. Association for Computational Linguistics.

Mathieu Ravaut, Aixin Sun, Nancy Chen, and Shafiq Joty. 2024. On context utilization in summarization with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2764–2781.

Sangwon Ryu, Heejin Do, Yunsu Kim, Gary Lee, and Jungseul Ok. 2024. Multi-dimensional optimization for text summarization via reinforcement learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5858–5871.

Brendan Walsh, Mark Hamilton, Greg Newby, Xi Wang, Serena Ruan, Sheng Zhao, Lei He, Shaofei Zhang, Eric Dettinger, William T. Freeman, and Markus Weimer. 2023. Large-scale automatic audiobook creation. In *Interspeech 2023*, pages 3675–3676.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zekun Xi, Wenbiao Yin, Jizhan Fang, Jialong Wu, Runnan Fang, Ningyu Zhang, Jiang Yong, Pengjun Xie, Fei Huang, and Huajun Chen. 2025. Omnithink: Expanding knowledge boundaries in machine writing through thinking. *arXiv preprint arXiv:2501.09751*.

A System Evaluation

A.1 Time Spent of User Evaluation Dataset

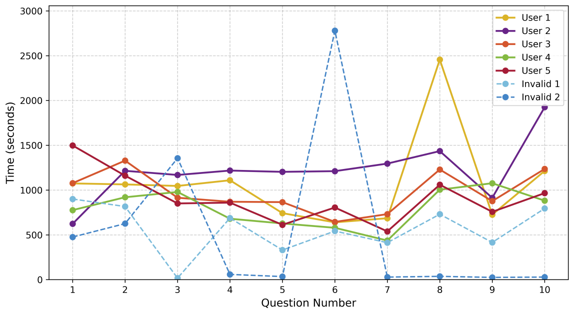
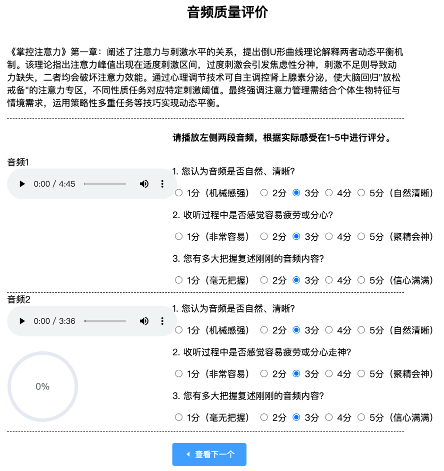


Figure 3: The time spent by the 7 evaluators on each element of the evaluation, with invalid 1 and invalid 2 being the users who discarded the results.

The evaluation time for each user is shown in Figure 3. The reading time of two users was very short, so they were considered invalid users, and their evaluation results were deemed invalid.

A.2 System Interface for Audio Evaluation



音频质量评价

《掌控注意力》第一章：阐述了注意力与刺激水平的关系，提出倒U形曲线理论解释两者动态平衡机制。该理论指出注意力峰值出现在适度刺激区间，过度刺激会引发焦虑性分神，刺激不足则导致动力缺失，二者均会破坏注意力效能。通过心理调节技术可自主调控肾上腺素分泌，使大脑回归"放松戒备"的注意力专区，不同性质任务对应特定刺激阈值。最终强调注意力管理需结合个体生物特征与情境需求，运用策略性多重任务等技巧实现动态平衡。

请播放左侧两段音频，根据实际感受在1~5中进行评分。

音频1

0:00 / 4:45

1. 您认为音频是否自然、清晰?

○ 1分（机械感强） ○ 2分 ● 3分 ○ 4分 ○ 5分（自然清晰）

2. 收听过程中是否感觉容易疲劳或分心?

○ 1分（非常容易） ○ 2分 ● 3分 ○ 4分 ○ 5分（聚精会神）

3. 您有多大把握复述刚刚的音频内容?

○ 1分（毫无把握） ○ 2分 ● 3分 ○ 4分 ○ 5分（信心满满）

音频2

0:00 / 3:36

0%

1. 您认为音频是否自然、清晰?

○ 1分（机械感强） ○ 2分 ● 3分 ○ 4分 ○ 5分（自然清晰）

2. 收听过程中是否感觉容易疲劳或分心走神?

○ 1分（非常容易） ○ 2分 ● 3分 ○ 4分 ○ 5分（聚精会神）

3. 您有多大把握复述刚刚的音频内容?

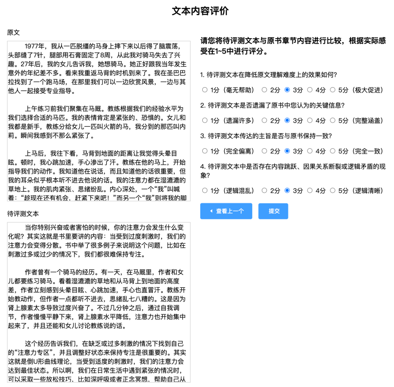
○ 1分（毫无把握） ○ 2分 ● 3分 ○ 4分 ○ 5分（信心满满）

Figure 4: Screenshot of audio evaluation.

As shown in Figure 4, users listened to randomly ordered audio clips from our system and FD, then completed a survey evaluating three aspects: (1) Naturalness—how fluent and natural the audio sounded, (2) Concentration—whether they felt fatigued or distracted, and (3) Comprehension—how well they understood the content.

A.3 System Interface for Interpretation Script Evaluation

As shown in Figure 5, users first read the original chapters to ensure a thorough understanding before evaluating the interpretation script. The assessment covered four dimensions: (1) Simplicity—how effectively the script reduced comprehension difficulty, (2) Completeness—whether key information was omitted, (3) Accuracy—consistency of main ideas with the original text, and (4) Coherence—absence of logical inconsistencies or contradictions.



文本内容评价

原文

1977年，我从一匹脱缰的马身上摔下来以后得了脑震荡，头部缝了7针，腿部用石膏固定了8周，从此我对骑马失去了兴趣。27年后，我的女儿告诉我，她想骑马。她正好跟我当年发生意外的年纪差不多。看来我重返马背的时机到来了。我在圣巴巴拉找到了一个跑马场，在那里我们可以一边欣赏风景，一边与其他人一起接受专业指导。

上午练习前我们聚集在马厩。教练根据我们的经验水平为我们选择合适的马匹。我的表情肯定是紧张的、恐惧的。女儿和我都是新手，教练分给女儿一匹叫火箭的马，我分到的那匹叫内莉。瞬间我感到不那么紧张了。

上马后，我往下看，马背到地面的距离让我觉得头晕目眩。顿时，我心跳加速，手心渗出了汗。教练在他的马上，开始指导我们的动作。我知道他在说话，而且知道他的话很重要，但我的耳朵似乎根本听不进去他说的话。我的注意力都在湿漉漉的草地上。我的肌肉紧张、思绪纷乱。内心深处，一个“我”叫喊着：“趁现在还有机会，赶紧下来吧！”而另一个“我”则将我的脚

待评测文本

当你特别兴奋或者害怕的时候，你的注意力会发生什么变化呢？其实这就是书里要讲的内容：当受到过度刺激时，我们的注意力会变得分散。书中举了很多例子来说明这个问题，比如在刺激过多或过少的情况下，我们都很难保持专注。

作者曾有一个骑马的经历。有一天，在马厩里，作者和女儿都要练习骑马。看着湿漉漉的草地和从马背上到地面的高度差，作者立刻感到头晕目眩、心跳加速，手心也直冒汗。教练开始教动作，但作者一点都听不进去，思绪乱七八糟的。这是因为肾上腺素太多导致过度兴奋了。不过几分钟之后，通过自我调节，作者慢慢平静下来，肾上腺素水平降低，注意力也开始集中起来了，并且还能和女儿讨论教练说的话。

这个经历告诉我们，在缺乏或过多刺激的情况下找到自己的“注意力专区”，并且调整好状态来保持专注是很重要的。其实这就是倒U形曲线理论，当受到适度的刺激时，我们的注意力会达到最佳状态。所以啊，我们在日常生活中遇到紧张的情况时，可以采取一些放松技巧，比如深呼吸或者正念冥想，帮助自己从

请您将待评测文本与原书章节内容进行比较，根据实际感受在1~5中进行评分。

1. 待评测文本在降低原文理解难度上的效果如何?

○ 1分（毫无帮助） ○ 2分 ● 3分 ○ 4分 ○ 5分（极大促进）

2. 待评测文本是否遗漏了原书中您认为的关键信息?

○ 1分（遗漏许多） ○ 2分 ● 3分 ○ 4分 ○ 5分（完整涵盖）

3. 待评测文本传达的主旨是否与原书保持一致?

○ 1分（完全偏离） ○ 2分 ● 3分 ○ 4分 ○ 5分（完全一致）

4. 待评测文本中是否存在内容跳跃、因果关系断裂或逻辑矛盾的现象?

○ 1分（逻辑混乱） ○ 2分 ● 3分 ○ 4分 ○ 5分（逻辑清晰）

Figure 5: Screenshot of interpretation script evaluation.

B System Interface

We have designed a concise, intuitive user interface[9] to optimize the user experience, as illustrated in Figure 6. The homepage (A) displays a list of books processed by the system. By clicking on a book cover, users can directly access the Audiobooks page (B) to access audio interpretations and related mind maps. This straightforward design enables users to quickly locate their desired books while significantly reducing operational complexity.

On the audiobook page, we offer two interpretation modes: full-book interpretation and chapter-by-chapter interpretation. Users can browse the audio list for a book and listen to the corresponding content by clicking the play button. Each audio entry is clearly labeled with the title, duration, and author information, allowing users to select specific chapters based on their needs. Additionally, the interface includes practical features like bookmarking, sharing, and subscribing to enhance usability and interactivity.

By combining auditory and visual sensory experiences, our design provides high-quality audio interpretations while leveraging mindmap to help

[9] If the HTTPS URL is inaccessible, you may try the HTTP URL as an alternative: http://49.232.199.229:14558/

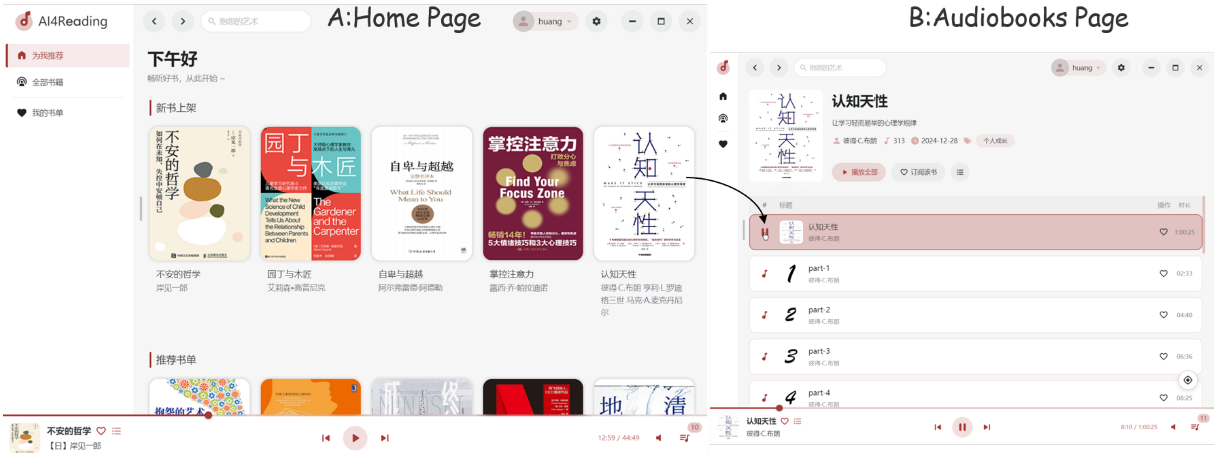


Figure 6: Screenshots of system interface.

users intuitively organize the core content and logical structure of the books. This multimodal learning approach enhances users' understanding of the material, improving both learning efficiency and overall reading experience. Whether for fragmented learning or systematic reading, the interface caters to diverse user needs, providing an immersive learning experience.