RAM3C: Enhancing Goal-oriented Open-ended Dialogue-based Educational System by Retrieval-augmented Multi-role agents Collaboration

Anonymous ACL submission

Abstract

This study presents the Retrieval-Augmented Multi-role Multi-agent Multi-round Collaboration (RAM3C) system, designed to improve the overall effectiveness of open-ended dialogue-based educational systems. Focusing on aspects of **H**umanlikeness, **I**ndividualization, **T**eaching expertise and **S**afety (**HITS**), RAM3C utilizes a dynamic framework that incorporates multi-agent, multi-round collaboration with multiple roles to harness collective expertise. RAM3C equips agents with tailored, multi-source knowledge bases and implements a history-sampling weighted retrieval-fusion approach to generate diverse, accurate, and safe educational dialogues. Our evaluation on a scenario of "Literature Discussion Class" by human volunteers and a decentralized, LLM-emulated expert group, confirms RAM3C's capability to deliver high-quality educational experiences, underscoring its substantial potential to elevate educational quality.

1 Introduction

Dialogue-based intelligent educational systems (DIES) execute "AI-to-student" dialogues to provide individualized educational services, which are significant for improving educational equity and enhancing the quality of education. Existing systems mainly teach given exercises and judge whether a student has mastered a piece of knowledge. However, the emergence of large language models (LLMs) has driven the development of DIES towards open-ended dialogue systems oriented to high-level educational goals, such as enabling students to grasp the core ideas of literary works, improving reading and critical thinking skills. This requires LLMs to possess comprehensive capabilities, including logical reasoning and knowledge answering, to meet four dimensions simultaneously: **H**umanlikeness, **I**ndividualization, **T**eaching expertise, and **S**afety, abbreviated as **HITS** (Detailed definition can be found in Appx.A).

To meet these requirements, there are challenges such as: **1)** For educational-goal oriented open-ended dialogue (EGOOD) tasks, LLMs' internal knowledge may be incomplete, inaccurate, or outdated, and often not aligned with professional educational theories or standards. **2)** There are few available educational cases, and hand-crafted prompts by educational experts can hardly cover open dialogues in versatile scenarios comprehensively, with high labor costs. **3)** High-quality fine-tuning data is unavailable, hindering the realization of individualization and humanlikeness through supervised fine-tuning. Therefore, a single LLM is hard to be competent for EGOOD Task.

To address these challenges, we propose the **R**etrieval-**A**ugmented **M**ulti-role **M**ulti-agent **M**ulti-round **C**ollaborative dialogue system (RAM3C). RAM3C dynamically organizes the collaborative process of multi-agents with various roles, progressively revising the original content generated by LLMs to meet **HITS**. Each round of revision is completed by the collaboration of multi-agents with a specific role. Each agent is equipped with different customized external knowledge bases, including multi-source data from teaching recordings to educational theories, and provides multi-scale accurate domain knowledge through the self-reflective RAG-Fusion technology, inserted into prompts as dynamic few-shot demonstrations, to improve the generation on **HITS**. We conducted experiments in the "Literature Discussion Class" scenario, dynamically identifying and constructing various roles such as "Language Education Expert", "Educational Psychologist", "Classic Author", "Cultural Safety Expert", and "Peer Learner". We had GPT-4 play the role of primary school students with various characteristics to generate dialogue records, and organized human volunteers for the experience. After the evaluation by GPT-4 and human vol-

unteers, our system achieved high-quality open educational dialogues.

2 Related Work

Dialogue-based educational systems, focusing on individualized guidance(Chen et al., 2023) and educational resource optimization(Deng et al., 2023), have been thoroughly explored. These systems, often powered by LLMs, play a supportive role by delivering exercises, recommending resources, and tracking student progress(Dan et al., 2023). Despite their contributions, they typically feature limited dialogue openness(Macina et al., 2023) and have not extensively addressed the complex challenges of higher-level educational goals which face the challenge of HITS(Kuhail et al., 2023).

The capability for reasoning and knowledge-based QA are core competences of LLMs in handling open-ended educational dialogue tasks (Long et al., 2024). The reasoning ability has been significantly enhanced through prompt engineering techniques. By assigning different role profiles to LLMs, displaying reasoning paths and examples, LLMs perform well in reasoning-intensive tasks (Wei et al., 2022; Besta et al., 2023; Wang et al., 2023b, 2022), role-playing tasks (Wang et al., 2023a; Zhou et al., 2023; Lu et al., 2024), knowledge QA tasks (Tang et al., 2023; Nori et al., 2023), and even creative tasks (Zhao et al., 2024). Meanwhile, retrieval-augmented generation (RAG) techniques like Self-RAG Asai et al. (2023) and query rewriting (Ma et al., 2023) improve LLMs' generation accuracy on knowledge-intensive tasks (Gao et al., 2023) by providing high-quality external knowledge to the input prompt.Despite rapid progress in these well-defined tasks, LLMs face the multidimensional **HITS** challenges in broader EGOOD tasks.

3 Methodology

This section introduces the design of RAM3C framework, as shown in Fig.1. RAM3C includes three procedures: 1) multi-role agents gathering, 2) retrieval-augmented single agent generation, and 3) multi-round multi-agent collaboration.

3.1 Multi-role agents gathering

The gathering of multi-role agents consists of the initialization of fixed-role agents and the generation of dynamic-role agents. Fixed-role agents participate in every round, including **Chinese language teachers**, **Educational psychologists** and **Culture safety experts**, who are responsible for the teaching expertise, humanlikeness, and safety of the generated content, respectively. Dynamic-role agents are gathered during the dialogue, to give specific advice. For example, the virtual literature author will be generated if Given the collection of students' speeches in round t, $\mathrm{SR}_t = \{sr_{1,t}, sr_{2,t} \ldots, sr_{N_{\text{stu}},t}\}$, where N_{stu} is the number of students in the discussion,

$$\texttt{prompt}_{\texttt{dr}}, \texttt{sys}_{\texttt{dr}} = \mathrm{LLM}(\texttt{prompt}_{\texttt{dyna}}, \texttt{sys}_{\texttt{dyna}}, \mathrm{SR}_t, \texttt{class_profile}) \quad (1)$$

where $\texttt{prompt}_{\texttt{dr}}$ and $\texttt{sys}_{\texttt{dr}}$ represent the guideline prompt and system role prompt of the newly generated dynamic role, respectively.

3.2 Retrieval-augmented single agent generation

Query rewriting. Similar queries are generated by agents before the retrieval. Original query OQ is the concatenation of students' responses SR and original response r generated by LLM to be modified.

$$\mathrm{SQ} = \mathrm{LLM}(\texttt{prompt}_{\texttt{fusion}}, \texttt{sys}_{\texttt{fusion}}, \mathrm{SR}||\mathrm{OQ}) \quad (2)$$

Historical sampling weighted reciprocal rank fusion. We propose Historical sampling weighted reciprocal rank fusion (HSW-RRF) algorithm to retrieve individualized reference for different students and agents. Specifically, for a given student, RAM3C maintains the historical sampling number n_d^{sample} and sampling frequency freq_d for each document d in the above knowledge base D. After retrieving, n_d^{sample} will be updated if retrieved. And freq_d will be updated as below:

$$\mathrm{freq}_d = \frac{n_d^{\text{sample}}}{\sum_{i \in D} n_i^{\text{sample}}} \quad (3)$$

Therefore the ranking weight W_d^{freq} can be updated as follows:

$$W_d^{\text{freq}} = \frac{e^{\mathrm{freq}_d}}{\sum_{i \in D} e^{\mathrm{freq}_q}} \quad (4)$$

HSW-RRF score can be calculated as

$$\mathrm{Score}(d \in D^{\text{sample}}) = \sum_{q \in Q} \frac{W_d^{\text{freq}}}{k + q(d)} \quad (5)$$

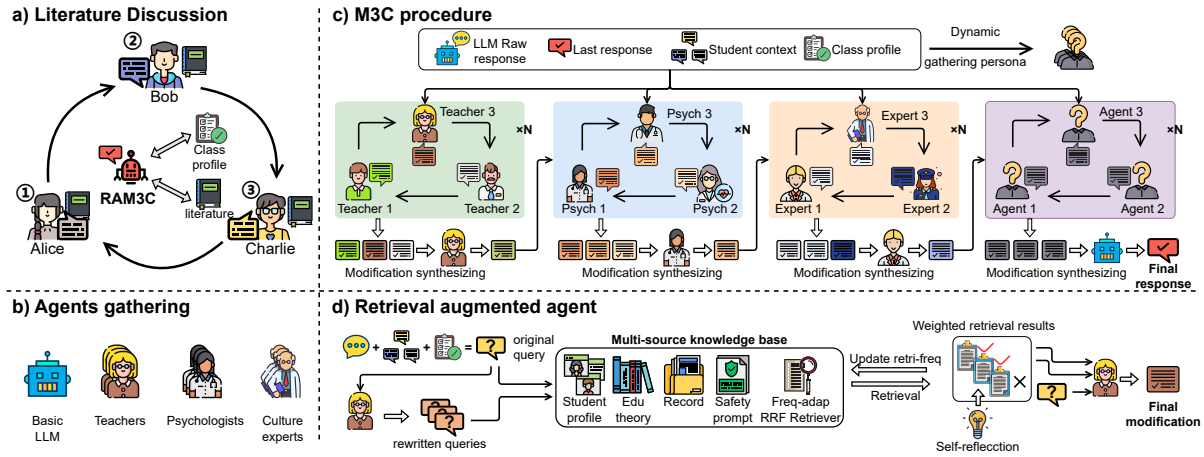

Figure 1: The schema of RAM3C. **a)** The schema of the experimental scenario, Literature Discussion Class, where 2-4 students discuss literature works under the guidance of the teacher. **b)** RAM3C gather fixed-role agents, basic LLM, Chinese language teachers, educational psychologists and culture safety experts, before the discussion. **c)** Multi-role agents collaboration. The final response of LLM is sequentially modified by groups of different experts according to the response in the last round, students' speeches, and the class profile, which contains the detail settings of this class. **d)** The design of the retrieval-augmented agent. Agents in **c)** are equipped with the customized external knowledge base, where diverse knowledge is retrieved by the proposed HSW-RRF retriever. The retrieval results are self-reflected by the agent. The acceptance (or not) of the raw retrieval results is used to update the history sampling frequency.

where D^{sample} is the set of retrieved documents by the query set Q, $k = 60$ is the constant from original RRF design (Cormack et al., 2009), $q(d)$ is the ranking of document d among all the documents of query q.

Agent generation. Under the guidance of retrieved documents D_i^{sample} and students' context SR_t, the agent i is able to analyze the original response res_{LLM} to be modified and given a professional response analysis ra:

$$\begin{aligned}\text{ra}_i^{\text{role}} =& \text{LLM}(\texttt{prompt}_{\texttt{role}}, \texttt{sys}_{\texttt{role}}, \\ & \text{res}_{\text{LLM}}, \text{SR}_t, D_i^{\text{sample}})\end{aligned} \tag{6}$$

where $\texttt{role} \in \{$teacher, psychologist, culture_expert, dyna_role$\}$.

3.3 Multi-role agents collaboration

After synthesizing the generation of each agent group of all roles, i.e.

$$\begin{aligned}\text{ra}^{\text{role}} =& \text{LLM}(\texttt{prompt}_{\texttt{role_gather}}, \\ & \texttt{sys}_{\texttt{role}}, \{\text{ra}_i^{\text{role}}\}_{i=1}^{N_{\text{agent}}}, \text{SR}_t)\end{aligned} \tag{7}$$

into the group modification ra^{role} of this agent group, RAM3C is then able to summarize the final generation $\text{ra}_t^{\text{final}}$ at round t by analyze $\{\text{ra}^{\text{role}}\}_{\text{role=teacher}}^{\text{dyna_role}}$.

4 Experiments

We verifies the proposed RAM3C's capability to generate high-quality content for educational oriented open-ended dialogue tasks.

Scenario setup. We select the scenario of "Literature Discussion Class" as an demonstration of an open-ended dialogue educational system. In this scenario, students discuss several topics about "Robinson Crusoe" under the guidance of a teacher who provides real-time feedback to promote the progress of dialogue-based teaching. Each expert group is equipped with 3 expert agents to ensure the dialogue content's expertise and diversity.

Automatic generation of dialogue topics. In the experiment, dialogue topics are automatically generated before the class. We design topics of five difficulty levels and randomly generate 10 topics (or questions) at a difficulty ratio of 1:3:3:2:1. Topics are randomly sampled from seven categories. The details of difficulty levels and topic types can be found in Appx.B.

Basic LLM and LLM-emulated students.. We use API of GLM-4[1], GPT-3.5, and GPT-4[2] as the basic LLMs and also the models with strong Chi-

[1] open.bigmodel.cn
[2] platform.openai.com/docs/models

Criteria \ Model		GPT-3.5 Turbo	GPT-4 turbo	GLM-4	Baichuan2-13B-Chat	Qwen1.5-14B-Chat
Humanlikeness	Emotional Feedback	8.4 ± 0.2	7.6 ± 0.7	8.0 ± 0.1	7.7 ± 0.2	8.7 ± 0.8
Expertise	Literary Understanding	8.9 ± 0.5	7.5 ± 0.5	9.3 ± 0.3	9.0 ± 0.8	8.9 ± 0.6
	Accurate Memory & Response	9.3 ± 0.3	8.8 ± 0.4	9.5 ± 0.2	7.9 ± 0.7	9.1 ± 0.1
	Educational Standard	7.2 ± 0.4	7.5 ± 0.2	8.8 ± 0.1	8.1 ± 0.6	6.9 ± 0.5
Individualization	Adaptive Dialogue	7.6 ± 0.2	7.4 ± 0.1	7.4 ± 0.4	9.2 ± 0.4	8.5 ± 0.3

Table 1: The evaluation score on humanlikeness, teaching expertise and individualization, graded by a GPT-4 expert group. The average and standard deviation of scores are calculated on 10 simulated dialogues, each of which contains 10 dialogue topics.

nese dialogue capabilities, such as Baichuan2-13B-Chat(Baichuan, 2023) and Qwen1.5-14B-Chat(Bai et al., 2023). To generate dialogue data, we configure LLM-emulated students to interact with the RAM3C, simulating three distinct fifth-grade Chinese primary school students, including two boys and one girl. The configurations are detailed in the Appx.E.

Multi-source knowledge base. Based on LangChain[3], Chromadb[4], and the embedding model BGE-m3(Chen et al., 2024), we build a multi-source knowledge base containing six types of data/knowledge. Details in Appx.C.

Hybrid peer-evaluation. To enhance dialogue evaluation efficiency and mitigate biases inherent in single LLMs, which may favor certain content, lengths, or styles(Liu et al., 2023), we adopt a combination of LLM experts to assess the quality of dialogues. Therefore, we convene a hybrid group of experts, including Chinese language teachers, educational psychologists, and cultural safety experts, to conduct decentralized peer evaluation. Each agent evaluates the entire dialogue independently. Then, agents from different roles score the evaluation opinions of other agents. Finally, the final evaluation result is obtained by synthesizing individual expert opinions and "expert-to-expert" scores. Hybrid peer evaluation leverages the collective expertise and diverse perspectives of all experts, ensuring the objectivity of the assessment.

Dialogue evaluation. Using the hybrid peer-evaluation above, we evaluate the simulated dialogue content according to the **HITS**, focusing on humanlikeness, individualization, and teaching expertise on 10 generated dialogues in Tab.1. We also invite 5 human volunteers to evaluate the models in Tab.3. By comparing the evaluation of GPT-4 and human, the GPT-4 is more optimistic than human experts, tending to assign higher scores. In contrast, human volunteers are more cautious in their scoring, especially on the items of "Accurate Memory and Response". However, the relative scoring by the GPT-4 expert group and human evaluators is consistent, indicating the effectiveness of LLM-based evaluation and the usability of the overall performance of RAM3C.

5 Conclusion

The RAM3C system enhances educational dialogue system on HITS requirements by leveraging the retrieval-augmented multi-role agents collaboration. Despite its promise, it faces limitations like incomplete knowledge base coverage and challenges in large dialogue managements. Future efforts will aim to expand knowledge integration and improve dialogue handling. In shorts, this work highlights AI's potential in education while recognizing the need for continued improvement.

6 Limitations

RAM3C may face negative impact on specific students or specific speech caused by bias from an external knowledge base. However, the screening and filtering of high-quality external knowledge may limit the scalability and underlying security of the system.

[3] https://github.com/langchain-ai/langchain
[4] https://github.com/chroma-core/chroma

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, et al. 2023. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

Yulin Chen, Ning Ding, Hai-Tao Zheng, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Empowering private tutoring by chaining large language models. *arXiv preprint arXiv:2309.08112*.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.

Yuhao Dan, Zhikai Lei, Yiyang Gu, Yong Li, Jianghao Yin, Jiaju Lin, Linhao Ye, Zhiyan Tie, Yougen Zhou, Yilei Wang, et al. 2023. Educhat: A large-scale language model-based chatbot system for intelligent education. *arXiv preprint arXiv:2308.02773*.

Yang Deng, Zifeng Ren, An Zhang, Wenqiang Lei, and Tat-Seng Chua. 2023. Towards goal-oriented intelligent tutoring systems in online education. *arXiv preprint arXiv:2312.10053*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Mohammad Amin Kuhail, Nazik Alturki, Salwa Alramlawi, and Kholood Alhejori. 2023. Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, 28(1):973–1018.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yun Long, Haifeng Luo, and Yu Zhang. 2024. Evaluating large language models in analysing classroom dialogue. *arXiv preprint arXiv:2402.02380*.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. *arXiv preprint arXiv:2401.12474*.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models. *arXiv preprint arXiv:2305.14283*.

Jakub Macina, Nico Daheim, Lingzhi Wang, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Opportunities and challenges in neural dialog tutoring. *arXiv preprint arXiv:2301.09919*.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

Hao Sun, Zhexin Zhang, Jiawen Deng, Jiale Cheng, and Minlie Huang. 2023. Safety assessment of chinese large language models. *arXiv preprint arXiv:2304.10436*.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023a. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023b. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona selfcollaboration. *arXiv preprint arXiv:2307.05300*, 1(2):3.

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.

Yunpu Zhao, Rui Zhang, Wenyi Li, Di Huang, Jiaming Guo, Shaohui Peng, Yifan Hao, Yuanbo Wen, Xing Hu, Zidong Du, et al. 2024. Assessing and understanding creativity in large language models. *arXiv preprint arXiv:2401.12491*.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, et al. 2023. Characterglm: Customizing chinese conversational ai characters with large language models. *arXiv preprint arXiv:2311.16832*.

A HITS: multi-dimensional requirements for EGOOD Tasks

Unlike general reasoning-intensive and knowledge-intensive tasks, educational goal-oriented open-ended dialogue tasks require LLMs to possess comprehensive capabilities while meeting the following multi-dimensional requirements.

1. **Humanlikeness**
   (a) **Emotional Feedback**: Through interaction with students, LLM should be able to recognize students' emotional states and respond appropriately with emotional feedback, such as comfort, encouragement, or sharing joy, thus establishing a deeper emotional connection.
2. **Individualization**
   (a) **Adaptive dialogue**: LLMs need to adapt their communicative style to align with the students' age, knowledge levels, and interests.
   (b) **Learner Modeling**: LLM should be able to dynamically customize learning content, difficulty, and paths based on learning progress, student preferences, and historical interactions.
3. **Teaching Expertise**
   (a) **Literary Understanding**: LLMs should have an in-depth understanding of literary works, including their themes, characters, plots, and literary techniques.
   (b) **Accurate Memory and Response**: LLMs must possess a accurate knowledge base for delivering factually correct responses.
   (c) **Heuristic Dialogue**: LLMs should employ a heuristic teaching method, guiding students to think through questions and discussions rather than just providing answers.
   (d) **Compliance with Educational Standards**: LLMs should adhere to given educational standards and theories.
4. **Safety**
   - **Content Appropriateness**: LLMs must ensure their outputs are devoid of inappropriate language, violence, sexual content, or any other material that could negatively affect students.
   - **Attack Robustness**: LLMs should be able to resist malicious use or attacks, such as attempts to induce the model to output inappropriate information by inputting malicious content.
   - **Data Privacy**: LLMs need to ensure it does not collect, store, or disseminate students' personal information, educational data from schools, and should as much as possible base model training and inference on local data.

B Educational goal auto-generation

To achieve individualized education, educational objectives are automatically generated by a group of Chinese language experts based on the student profile and external hyper-parameters (in Table x). Taking the scenario of "Literature Discussion Class" as an example, a systematic educational objective consist of several discussion topics, each with different levels of difficulty and aiming at various related abilities. After the initial generation of topics, manual intervention (optional) and other expert groups are involved in modifications before the topics are finalized for use in the class. See Appx.B.1 and B.2 for the detail.

B.1 Category of dialogue topics

Dialogue topics are divided into seven categories, covering a wide range of dimensions related to Chinese language reading classes and the cultivation of reading abilities.

1. **Reading Comprehension**: The goal is to help students understand and interpret the content of texts, including the plot, characters, and setting. Teach students to identify the main ideas and supporting details of a book.
2. **Language Skills**: Improve students' oral and written expression through discussion and writing. Strengthen the use of grammar, vocabulary, and rhetoric.
3. **Cultural and Historical Awareness**: Provide cultural and historical knowledge through the background of classic literature. Enhance students' understanding of different eras and social contexts.

4. **Emotions and Values**: Guide students in emotional and values education through the discussion of moral and ethical issues in classic literature. Cultivate empathy and critical self-reflection.

5. **Critical Thinking**: Encourage students to conduct in-depth analysis and critically evaluate the viewpoints and arguments in the literature work. Develop students' ability to examine issues from multiple perspectives.

6. **Creative Thinking**: Inspire students' imagination and encourage them to think and express creatively. Enhance innovation skills through activities like rewriting plots or creating alternative endings.

7. **Integrated Learning Skills**: The teaching model should encourage students to integrate and apply multidisciplinary knowledge. Promote interdisciplinary thinking, such as linking literary works with history, sociology, or philosophy.

B.2 Difficulty level

In the scenario of Literature Discussion Class, the topics differ from five difficulty levels, corresponding to the reading-related abilities from low level to high level.

Level 1 **Knowledge and memory**: Direct questions about the basic plot or characters in the literature work. Example: 1) What was Robinson's main occupation in 'Robinson Crusoe'? 2) Where was Robinson stranded in the story? 3) Explain how Robinson built his own dwelling.

Level 2 **Understanding**: Simple explanations and summaries about the story of the literature. Example: 1) Describe the first challenge Robinson encountered on the island. 2) Explain the strategies Robinson used to survive on the deserted island.

Level 3 **Application**: Applying information or concepts from the literature's story to new situations. Example: 1) How would Robinson's life have been different if he had modern tools on the island? 2) Discuss how Robinson's survival skills on the deserted island demonstrate human adaptability and creativity. 3) Which part of 'Robinson Crusoe' shows Robinson's resource management skills? Please provide specific examples.

Level 4 **Analysis**: Analyzing elements of the story in the literature work, such as theme, symbolism, or character motives. Example: 1) Analyze what the deserted island symbolizes in 'Robinson Crusoe'. How does Robinson's experience reflect the social and cultural background of that time? 2) Analyze what the deserted island symbolizes in 'Robinson Crusoe'. How does it reflect Robinson's inner world and growth? 3) Discuss how Robinson's relationship with Friday demonstrates the views on 'civilization' and 'barbarism' of the society at that time.

Level 5 **Evaluation and Creation**: Evaluating the themes or character decisions in the story, or proposing new storylines and characters. Example: 1) Evaluate Robinson's treatment of the indigenous people on the island. Do you think his actions were justified? Why or why not? Or, imagine a new ending, and explain your choice. 2) If you had the opportunity to rewrite 'Robinson Crusoe', how would you reset Robinson's adventure journey? Please explain the reasons behind your choice and the expected theme changes. 3) Evaluate Robinson's actions and decisions on the deserted island. From the perspective of modern society, are these actions and decisions still considered justified and reasonable? Why or why not?

C Multi-source knowledge base

We establish a multi-source knowledge base to support the multi-role agents' collaboration. The knowledge base includes the following sources of knowledge:

1. **Class dialogue records.** Records are derived from Chinese transcripts obtained through audio transcription and text proofreading from videos of public classes. These records demonstrate different teaching styles and responses that adhere to educational standards. The translation of the part of the records can be found at Appx.F.

Source	Counts
Dialogue records	1,688,000 words
Educational theories	3,770,000 words
Literature works	207,800 words
Edu-psycho theories	2,672,000 words
Safety prompts	13,893,188 words
Encyclopedia	196,494 items

Table 2: Summary of counts in Chinese character across different knowledge sources.

2. **Theories and research papers on Chinese language teaching.** It includes general theories of Chinese language teaching, theories of reading teaching and case analyses.

3. **Theories and case analyses in educational psychology.**

4. **Safety prompts.** Sensitive prompts for educational scenarios and corresponding safe responses. We use GPT-4 to filter and rewrite seven types of malicious prompts and their appropriate responses from (Sun et al., 2023), including crimes and illegal activities, ethics and morality, insult, mental health, physical harm, privacy and property, unfairness and discrimination, for reference by cultural safety experts.

5. **Encyclopedic knowledge in Chinese.** Encyclopedic knowledge is sampled from Wikipedia[5] to provide accurate answers related to background knowledge.

6. **literature works in Chinese** These texts support discussions involving the original plots of literary works.

D Human evaluation and ablation studies

Besides the experiments in the mainbody, we also conduct a series of ablation studies, including the impact of different numbers of experts in the expert group on the dialogue effect, the effect of role addition or reduction on dialogue, and the utility of the retrieval enhancement module. The details of these studies will be released on this URL[6] subsequently.

[5]https://huggingface.co/datasets/bigscience-data/roots_zh-cn_wikipedia

[6]https://github.com/RAM3C/RAM3C

Criteria	GPT-3.5 Turbo	GPT-4 Turbo	GLM-4
Emotional Feedback	8.0 ± 0.4	8.5 ± 0.2	8.8 ± 0.7
Literary Understanding	9.3 ± 0.1	9.0 ± 0.2	9.5 ± 0.5
Accurate Memory & Response	7.2 ± 0.4	7.4 ± 0.3	7.5 ± 0.6
Education Standard	8.2 ± 0.5	8.5 ± 0.6	8.0 ± 0.3
Adaptive Dialogue	7.5 ± 0.6	7.7 ± 0.8	8.1 ± 0.5

Table 3: The evaluation of GP-3.5 Turbo, GPT-4 Turbo and GLM-4 by five human volunteers.

E Prompt templates

Prompt templates used in the above experiments are listed as below. Prompts are written in Chinese and translated into English. The direct use of English prompts may result in different experimental results than those in the paper. Original Chinese prompts can be found in https://github.com/RAM3C/RAM3C.

(1) $\texttt{prompt}_{\texttt{qg}}$ for the original generation of one dialogue topic:
<book>: `book_name`.
<Difficulty level>: `difficulty_level`.
<Educational goal>: `goal`.
<Generated questions>: `questions`.
According to <Educational goal> and <Difficulty level>, generate **one** question. Make sure not to duplicate <Generated questions>.
Example: If you were Robinson Crusoe, how would you manage your relationship with the indigenous people on the island? Please try to create a scenario different from the original story, and explain the moral and ethical considerations behind your choice:).

(2) $\texttt{sys}_{\texttt{psy}}$ for system role prompt of the psychologist agents:
As a professional educational psychologist who understands students' learning motivation, cognitive development, and emotional needs, you possess profound theoretical knowledge and practical experience. You are capable of understanding and addressing the psychological challenges and needs students encounter during their learning process.

Your goal is to revise and polish the <TEXT TO BE MODIFIED> to align with students' psychological and emotional needs, supporting their holistic development in literary studies.

(3) $\texttt{prompt}_{\texttt{psy}}$ for the response modification of a psychologist agent:
<Dialogue topic>: `topic`.
<Students speech>: `student_responses`.
<TEXT TO BE MODIFIED>: `sentence`.
<Reference>: `theory`.
Please follow these requirements:
1. Show emotional care, using warm and friendly language.
2. Based on the principles of educational psychology and <Reference>, assess the psychological and emotional impact of <TEXT TO BE MODIFIED> on students, ensuring the content is appropriate for their age and development stage. Pay special attention to the suitability of the language and the accuracy of emotional expression.
3. Analyze <Students speech> to determine the student's emotional state (anger / excitement / discouragement / sadness / happiness / anxiety...), and provide targeted feedback according to different states.
4. Analyze whether psychological and emotional issues are involved in <Dialogue topic>; if so, provide a professional response based on <Reference>. If not, skip this step.
5. Keep your modification plan concise, without exceeding the length of the <TEXT TO BE MODIFIED>, and avoid lengthy explanations. **Do not** explain your intentions for modification; directly produce the modification plan.
Please generate the <modification>:

(4) $\texttt{prompt}_{\texttt{psy_gather}}$ for synthesizing the modifications from every psychologist agent:
<Students speech>: `student_responses`.
<Expert Modifications>: `expert_answers`.
As a professional educational psychologist, please follow your professional knowledge with the assistance of <Expert Modifications>, follow the requirements below to provide a <Final Modification> to <Students speech>:
1.Carefully and comprehensively assess each of <Expert Modifications>, paying special attention to any differences and contradictions that may exist among different experts.
2. Conduct a comprehensive summary, analysis, and necessary refinement of the <Expert Modifications> to ensure the content's language affinity, educational value, and professional depth.
3. Address the core message of the <Students speech> to ensure the integrated text remains true to the original theme and intent.
4. The response must match the cognitive level and vocabulary of a fifth-grade elementary student in China, avoiding abstract terms and advanced concepts.
5. Use a lively and vivid language style suitable for elementary students. The integrated solution should be concise and not exceed the length of <Expert Modifications> too much.
6. Do not mention <Expert Modifications>; it is you answering the classmates! Generate a clear, coherent, professionally scrutinized <Final Modification> based on the above information.

(5) $\texttt{sys}_{\texttt{RAG-fusion}}$: You are a helpful assistant that generates multiple search queries based on a single input query.

(6) $\texttt{prompt}_{\texttt{RAG-fusion}}$: Generate multiple search queries related to: <original_query>.

(7) $\texttt{sys}_{\texttt{teacher}}$: You are a Chinese language education expert with a deep understanding of language teaching, proficient in reading comprehension, literary analysis, and writing skills. You are familiar with various literary genres, writing styles, and rhetorical techniques, and excel in designing language teaching activities and dialogue topics related to "Literature Discussion Class". You can guide students to deeply explore the themes, symbols, and underlying meanings of texts. With extensive teaching experience, you are capable of designing engaging open-ended questions.

(8) $\texttt{prompt}_{\texttt{teacher_q}}$ for modifying one dialogue topic (or question) of a teacher agent:
<Dialogue topic>: `topic`.
<Students speech>: `student_responses`.
<TEXT TO BE MODIFIED>: `sentence`.
<Reference>:
<Book content>: `book`.
<Class record>: `record`.
<Educational theory>: `theory`.
Based on your professional knowledge, following the requirements below, analyze how to revise and polish the <TEXT TO BE MODIFIED>:
1. If a question is raised in <Students speech>, you must first be succinctly answered based on <Reference>, not exceeding **100 words**. Add your answer to the beginning of the <Modification>.
2. Your <Modification> must be suitable for the cognitive level and vocabulary of fifth-grade primary school students in China, avoiding abstract vocabulary and advanced concepts.

3. Refer to relevant content in <Educational theory> to enhance the language quality, literary depth, and teaching effectiveness of the content.
4. Refer to the language style and teaching methods in <Class record> to use a lively and vivid language style for primary school students in the <TEXT TO BE MODIFIED>.
5. Your <Modification> should be concise and should not exceed the length of the <text to be revised> by too much.
Example:
<TEXT TO BE MODIFIED>: In "Robinson Crusoe," what difficulties and challenges did Robinson face in the story?
<Modification>: Classmates, in "Robinson Crusoe," Robinson encountered many difficulties and challenges, both in life and psychologically. Can you tell me what difficulties he faced?
<TEXT TO BE MODIFIED>: In "Robinson Crusoe," how did Robinson use his skills and creativity to protect himself from the dangers and threats on the island?
<Modification>: Boys and girls, we know that the deserted island where Robinson was, was not a safe place. But Robinson used his courage and wisdom to protect himself from the dangers and threats on the island. How do you think Robinson managed to do that?
<TEXT TO BE MODIFIED>: How did Robinson adapt to life on the deserted island?
<Modification>: Classmates, do you feel lonely when you are alone at home? Robinson's life on the island must have been very lonely, too. How did he adapt to living on the deserted island?
Do not answer the questions in <TEXT TO BE MODIFIED>.
Please generate the <Modification>:

(9) $\texttt{prompt}_{\texttt{teacher_q_gather}}$ for synthesizing modifications of one given original dialogue topic (or question) from every teacher agent:
<Dialogue topic>: `topic`.
<Students speech>: `student_responses`.
<TEXT TO BE MODIFIED>: `sentence`.
* Do not answer the questions in <TEXT TO BE MODIFIED>.**
<Expert Modifications>: `expert_answers`.
Based on your own expertise, following the requirements below, integrate <Expert Modifications> to make <Final Modification>:
1. Carefully and comprehensively evaluate each piece of <Expert Modifications>, paying special attention to differences and contradictions that may exist between different experts.
2. Based on <Expert Modifications>, conduct a comprehensive summary analysis and necessary revisions and modifying of the <TEXT TO BE MODIFIED> to ensure the content's language affinity, educative nature, and professional depth.
3. During the modification process, preserve the core information and educational goals of the <TEXT TO BE MODIFIED>, ensuring that the integrated text remains faithful to the original theme and intent.
4. Must be suitable for the cognitive level and vocabulary of fifth-grade Chinese primary school students, avoiding the use of abstract vocabulary and advanced concepts.
5. Use a lively and vivid language style suitable for primary school students. Keep the integration plan concise, ideally ask only one question, and should not exceed the length of the <TEXT TO BE MODIFIED> too much.
Please generate a clear, coherent, and professionally scrutinized <Final Modification>:

(10) $\texttt{prompt}_{\texttt{teacher_a}}$ for generating the response of the given students' speech from one teacher agent:
<Dialogue topic>: `topic`.
<Students speech>: `student_responses`.
<Reference>:
<Book content>: `book`.
<Class record>: `record`.
<Educational theory>: `theory`.
Based on the <Book content> and your own professional knowledge, respond to <Students speech> by generating <Response> according to the following requirements:
1. If a question is raised in <Students speech>, it must first be succinctly answered based on <Reference>, not exceeding 100 words.
2. It must be suitable for the cognitive level and vocabulary of fifth-grade Chinese primary school students, avoiding the use of abstract vocabulary and advanced concepts.
3. Refer to related content in <Educational theory> to enhance the language quality, literary depth, and teaching effectiveness of <Response>.
4. Refer to the language style and teaching methods in <Class record> to use lively and vivid primary school language styles in <Response>.
5. Keep your <Response> concise, not to exceed twice the length of <Students speech>.
Generate <Response>:

(11) $\texttt{prompt}_{\texttt{teacher_a_gather}}$ for synthesizing re-

sponses generated by teacher agents:
<Students speech>: student_responses.
<Expert generations>: expert_answers.
You are a Chinese language teacher in a Chinese primary school. During your class, students speak up. Based on your own professional knowledge and with the help of <Expert generations>, follow the requirements below to give the <Final answer>:
1. Carefully and comprehensively assess each <Expert generations>, paying special attention to differences and contradictions that may exist between different experts.
2. Conduct a comprehensive summary analysis of the <Expert generations> and make necessary modifications to ensure the content's language affinity, educative nature, and professional depth.
3. Respond to the core messages of <Students speech>, ensuring that the integrated text remains true to the original theme and intent.
4. Must match the cognitive level and vocabulary of fifth-grade Chinese primary school students, avoiding abstract vocabulary and advanced concepts. 5. Use a lively and vivid language style suitable for primary school students.
6. Keep the integration plan concise, not exceeding the length of <Expert generations>.
7. Do not mention expert opinions, it is you who are answering the students!
Please generate a clear, coherent, and professionally reviewed <Final answer>:

(12) $\texttt{sys}_{\texttt{llm_student}}$ for the system role prompt of LLM-emulated students:
Role: You are a 10-year-old <boy / girl> fifth grader in a Chinese primary school. You are <lively and cheerful / imaginative / sensitive and delicate / full of creativity / introverted and shy / curious / confident and independent / rigorous and earnest / compassionate / diligent and studious>.
Scenario: You have just finished reading <book> and are curious about <topic>. You are participating in a <book> themed discussion class organized by your Chinese language teacher and attended by several classmates.
You must follow these requirements:
1. Based on the questions and guidance provided by the teacher, express your thoughts and answers in simple children's language without being long-winded!!
2. Pose a small question in response to the conversation's progress to maintain its continuity and liveliness.
3. Your reply should not exceed 200 Chinese characters.

(13) $\texttt{prompt}_{\texttt{llm_student}}$ for LLM-emulated student agents:
<Dialogue topic>: topic.
<Reference>: book.
<Other students speech>: other_student_sentence
As a 10-year-old Chinese elementary school student, please answer the question based on the <Dialogue topic> and <Reference>:
1. Use simple language suitable for children to present your thoughts and answers, no lengthy discourses!!
2. Do not simply repeat what <Other students speech>; have your own independent thoughts.
3. Use a variety of sentences and structures, avoiding repetition of what <Other students speech>.
Your response should not exceed 200 Chinese characters.
Occasionally, pose a small question to keep the conversation going and lively.

(14) $\texttt{prompt}_{\texttt{llm_eval}}$ for LLM-emulated expert for the dialogue content evaluation:
<Dialogue topic>: topic.
<Class records>: log.
<Class profile>: profile.
<Evaluation Criteria>:
1. Accuracy: Assess the accuracy of the educational content provided by the system in terms of facts and knowledge.
2. Engagement: Examine how the system engages students in dialogue, including the frequency and depth of interaction.
3. Individualization: Evaluate whether the system can adjust personalized settings according to the needs and responses of different students.
4. Educational Quality: Assess the effectiveness of the system in enhancing students' knowledge, thinking abilities, and other aspects.
5. Humanlikeness: Examine whether the system's dialogue is natural, similar to the communication style of human teachers, and whether it can simulate human emotions and empathy.
6. Safety: Assess the appropriateness of the system's content, whether it meets educational standards, and avoids inappropriate or sensitive topics.
Based on <Dialogue topic>, <Class records>, <Relevant Information>, and <Evaluation Criteria>, conduct a quantitative evaluation of the course qual-

ity, provide a score **from 1 to 10** for each criterion, and give specific reasons for the evaluation and suggestions for improvement.

(15) $\texttt{prompt}_{\texttt{llm_eval_score}}$ for scoring the evaluation of other LLM-emulated experts:
<Another expert's evaluation>: `other_eval`.
Follow the <Evaluation Criteria> to grade the <Another expert's evaluation>.
<Evaluation Criteria>:
1. Accuracy: Assess the accuracy of the educational content provided by the system in terms of facts and knowledge.
2. Engagement: Examine how the system engages students in dialogue, including the frequency and depth of interaction.
3. Individualization: Evaluate whether the system can adjust personalized settings according to the needs and responses of different students.
4. Educational Quality: Assess the effectiveness of the system in enhancing students' knowledge, thinking abilities, and other aspects.
5. Humanlikeness: Examine whether the system's dialogue is natural, similar to the communication style of human teachers, and whether it can simulate human emotions and empathy.
6. Safety: Assess the appropriateness of the system's content, whether it meets educational standards, and avoids inappropriate or sensitive topics.
Based on <Dialogue topic>, <Class records>, <Relevant Information>, and <Evaluation Criteria>, conduct a quantitative evaluation of the course quality, provide a score **from 1 to 10** for each criterion, and give specific reasons for the evaluation and suggestions for improvement.

F Translation of classroom dialogue records

Below is an English translation of a portion of the classroom dialogue records written in Chinese, retaining as much of the original style as possible.

> **Teacher**: Oh, absolutely! I'm right there in the book, waiting for you. That's the special bond between an author and their readers. Authors always wait in their books, silently hoping we'll drop by. Opening a book is like walking into the author's world, promising a never-miss-out date. So today, let's go on a date with a classic—"Robinson Crusoe." Let's kick things off. What do you guys know about the author? You, tell me.
>
> **Student1**: The author's Defoe, from England. Born in 1660, died in 1731. He wrote "Colonel Jack" and "Memoirs of a Cavalier" among other things.
>
> **Teacher**: Nice! You really know your stuff, here's three points for you! Who's next? Don't be shy, your turn.
>
> **Student2**: Oh, and Defoe got the idea for this book from a real story that happened over 200 years ago. This Scottish sailor had a fallout with his captain and ended up on a deserted island, living there for 4 years. That's what kicked off "Robinson Crusoe."
>
> **Teacher**: Look at you, another smarty! Points for you too. Anyone else? Go ahead.
>
> **Student3**: Just to add, the Robinson Crusoe Defoe wrote about was way different from that English sailor; Crusoe was all heroic and stuff.
>
> **Teacher**: And the real sailor, not so much with the heroic acts, huh? Great adding that. Points for you. Looks like you guys are not just good at digging up info but also great at putting it all together. That's an awesome skill to have!