Cognitive Strategies in Study
WISDOM ACADEMIC https://doi.org/10.63808/css.v2i3.426 Volume 2 , Issue 3

Research Article

Research on Personalized Popular Science Q&A Generation Method Integrating Difficulty-Aware Retrieval-Augmented Generation

Yuewen Cao[1,*], Xinyang Li[1], Chunyu Lei[1], Lin Zhang[1]
[1]School of Computer and Artificial Intelligence, Beijing Technology and Business University, Beijing 102488, China
[*]**Correspondence:** Yuewen Cao, 15313187326@163.com

Abstract

General large language models suffer from factual hallucinations, mismatched expression difficulty and insufficient individual adaptation in science education question-and-answer (Q&A) tasks. To address these issues, this paper proposes a personalized popular science Q&A generation method integrating difficulty-aware retrieval-augmented generation. Firstly, a scientific knowledge base with academic stage difficulty labels is constructed in this method. Secondly, learner portrait matching items and knowledge difficulty deviation penalty items are introduced into the traditional semantic retrieval score to realize evidence screening for students in different academic stages. In the generation stage, structured prompts are organized with retrieved evidence, student portraits and answer constraints, and parameter-efficient fine-tuning is adopted to enhance the model's adaptability to popular science Q&A style and knowledge boundaries. Taking a self-built primary and secondary school science Q&A dataset as the research object, experiments are conducted from the dimensions of scientific accuracy, age appropriateness, completeness and hallucination rate. The experimental results show that the proposed method outperforms conventional retrieval-augmented methods in terms of scientific accuracy and age appropriateness, and can improve the reliability and teaching applicability of popular science Q&A with low training costs.

Keywords

Science education, Large language model, Retrieval-augmented generation, Personalized Q&A, Parameter-efficient fine-tuning

1. Introduction

Science education emphasizes the collaborative cultivation of scientific concepts, inquiry

processes and evidence awareness. With the continuous expansion of the application of large language models in natural language generation (Kasneci et al., 2023; OpenAI, 2023), knowledge Q&A and teaching assistance, automated popular science Q&A for primary and secondary school scenarios has become an important research direction in intelligent education. Different from general open-domain Q&A, popular science Q&A requires not only accurate scientific facts, but also expressions that conform to learners' age, knowledge foundation and cognitive characteristics. For example, the explanation of "why solar eclipses occur" should focus on life-oriented analogies and intuitive interpretations for primary school students, while basic concepts such as linear propagation of light, lunar revolution and relative celestial position relationships can be further introduced for junior high school students.

Although general large language models have strong language organization capabilities, they still have three major problems when directly applied to science education. First, the models may generate unsubstantiated or factually incorrect content, namely the "hallucination" problem (Mallen et al., 2023). Second, the models tend to provide unified answers for all questions, failing to adaptively adjust the explanation depth and terminology density. Third, the models lack perception of textbook knowledge systems and academic stage differences, resulting in significant differences in the teaching applicability of the same answer for different students. Retrieval-Augmented Generation (RAG) improves factual reliability by introducing external knowledge bases (Lewis et al., 2020; Gao et al., 2023), but traditional RAG only focuses on the semantic similarity between questions and documents, ignoring the unique academic stage difficulty matching requirements in educational scenarios.

From the perspective of educational psychology, difficulty matching in popular science Q&A is not merely a matter of simplifying wording. It should place the answer within a cognitive range that learners can understand and further master with appropriate support. The Zone of Proximal Development emphasizes that instructional support should bridge learners' current and potential development levels (Vygotsky, 1978), while Cognitive Load Theory suggests that excessive terminology density and reasoning span may increase extraneous cognitive load and impair comprehension (Sweller, 1988). Therefore, this study incorporates academic-stage labels, difficulty deviation and learner portraits into retrieval and generation to strengthen the educational foundation of the proposed method.

Accordingly, this paper proposes a personalized popular science Q&A generation method integrating difficulty-aware retrieval augmentation. The core idea of the proposed method is to comprehensively consider question relevance, student portrait matching degree and knowledge fragment difficulty deviation in the retrieval stage; in the generation stage, structured prompts are used to constrain the model to answer based on evidence, and LoRA

parameter-efficient fine-tuning is adopted to enhance popular science expression adaptation. Different from large-scale training or complete reinforcement learning, the proposed method adopts a lightweight design, which is easy to implement under the experimental conditions of ordinary university projects and general journal papers. The main contributions of this paper are summarized as follows: (1) Construct a difficulty-labeled scientific education knowledge base and a student portrait representation method; (2) Propose a difficulty-aware retrieval scoring function to jointly optimize relevance and age appropriateness; (3) Design an evidence-constrained popular science answer generation process, and verify the effectiveness of the proposed method through comparative experiments and ablation experiments.

2. Related Work

2.1. Large Language Models and Popular Science Q&A

Large-scale pre-trained language models learn linguistic patterns and partial factual knowledge from massive corpora, and exhibit outstanding performance in open-domain Q&A, text summarization, educational assistance and other tasks. GPT-3 proposed by Brown et al. (2020) demonstrates the generalization ability of large models in few-shot learning scenarios. However, the knowledge embedded in pre-trained parameters cannot be directly tracked or updated, which may lead to errors in the interpretation of changing scientific facts, long-tail knowledge and professional concepts. Chain-of-thought prompting proposed by Wei et al. (2022) improves the performance of complex reasoning tasks, but for science education, prompts alone cannot solve the problems of factual traceability and academic stage adaptation.

In educational applications, model outputs are required to be not only “accurate” but also “learning-friendly”. Existing intelligent teaching systems usually improve learning effects through learner modeling, resource recommendation and personalized feedback. However, after integration with large language models, how to explicitly model learner portraits in the generation process remains a worthy research issue. This paper takes student portraits as joint constraints for retrieval and generation, and realizes low-cost personalization in a weakly supervised manner.

2.2. Retrieval-Augmented Generation

RAG combines parametric language models with non-parametric external knowledge bases, which can improve the factuality and interpretability of knowledge-intensive tasks(Lewis et al., 2022).Dense retrieval methods for open-domain Q&A map questions and documents to the same vector space through dual encoders, overcoming the literal

coincidence dependence of traditional BM25(Robertson & Zaragoza, 2009) algorithms (Karpukhin et al., 2020). In recent years, retrieval-augmented methods have been widely applied in Q&A, summarization, code generation, domain knowledge services and other tasks(Gao et al., 2023).

Although RAG can reduce the risk of model hallucinations(Gao et al., 2023; Mallen et al., 2023), the retrieval goal of conventional RAG is to obtain the "most relevant" content rather than the "most suitable" content. In science education scenarios, retrieved evidence containing excessive beyond-curriculum terminology and abstract derivation will reduce the intelligibility of answers. Therefore, this paper incorporates learner portraits and knowledge difficulty factors into the retrieval scoring function, enabling the retrieval process to meet both factual support and age-appropriate expression requirements.

2.3. Parameter-Efficient Fine-Tuning and Personalized Adaptation

Full-parameter fine-tuning of large language models requires high computing power and storage costs, which is not suitable for lightweight educational applications. LoRA freezes pre-trained weights and injects low-rank trainable matrices to realize task adaptation with a small number of parameters(Hu et al., 2022). Parameter-efficient fine-tuning methods such as Prefix-tuning and Adapter also provide feasible paths for domain migration(Li & Liang, 2021; Houlsby et al., 2019). This paper adopts LoRA as a lightweight adaptation method, aiming to make the model stably follow popular science answer specifications with limited samples, rather than training a brand-new educational large language model.

2.4. Educational Psychology Foundation

In the learning sciences, the Zone of Proximal Development argues that learning materials should be slightly above what learners can complete independently, but still understandable with external support(Vygotsky, 1978). This provides a theoretical basis for the difficulty-deviation constraint in this paper: retrieved evidence should neither be too simple to be pedagogically meaningful nor far beyond the learner's current academic stage. Cognitive Load Theory further indicates that learners' working memory is limited, and excessive irrelevant terminology, complex symbols and long reasoning chains may increase extraneous cognitive load(Sweller, 1988). Accordingly, the proposed method penalizes academic-stage difficulty deviation during retrieval and controls terminology density and explanation depth during generation, so that the output better matches students' cognitive processing capacity.

3. Proposed Difficulty-Aware Retrieval-Augmented Q&A Generation

WISDOM ACADEMIC https://doi.org/10.63808/css.v2i3.426 Volume 2 , Issue 3

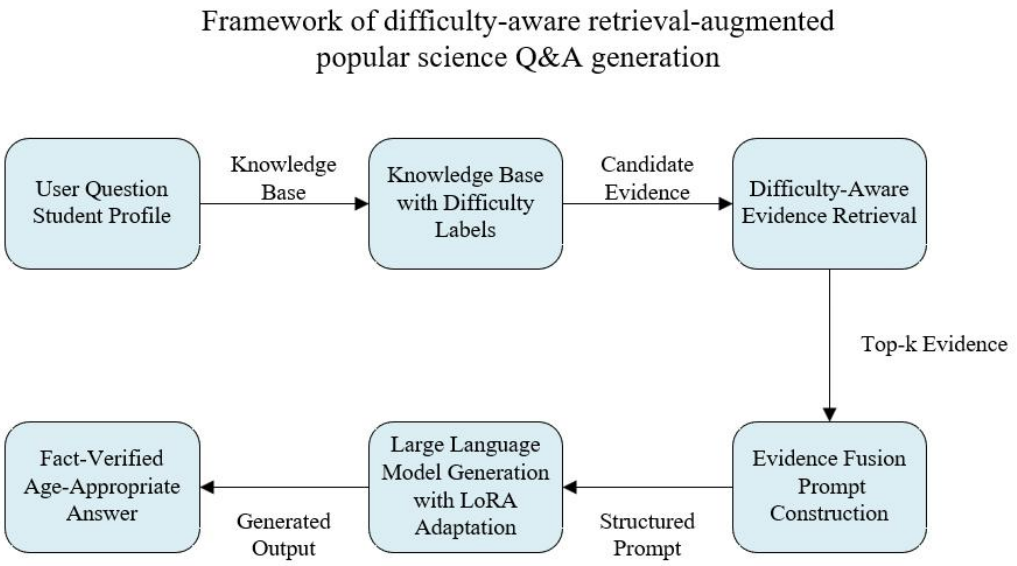
Method

To mitigate hallucinations, difficulty mismatch and insufficient personalization in science Q&A, we present a difficulty-aware retrieval-augmented generation framework. As depicted in Figure 1, it comprises difficulty-aware retrieval and evidence-constrained generation.

The retrieval module computes semantic similarity between the question and knowledge fragments, then re-ranks candidates by integrating learner portrait matching and difficulty deviation penalty to select age-appropriate evidence. The generation module constructs structured prompts with evidence, portrait and constraints, and adapts a pre-trained LLM via LoRA to produce accurate, age-appropriate answers.

Figure 1

Framework of difficulty-aware retrieval-augmented popular science Q&A generation



Note. The overall framework includes two core modules: a difficulty-aware evidence retrieval module and an evidence-constrained answer generation module with LoRA parameter-efficient adaptation.

3.1. Problem Definition

Given a student's scientific question q, student portrait u, and scientific knowledge base $D=\{d_1,d_2,...,d_n\}$, the goal is to generate an answer y that satisfies scientific accuracy, age-appropriate expression and content completeness simultaneously. The student portrait u consists of academic stage g_u, knowledge foundation b_u, expression preference s_u and other attributes. Each knowledge fragment d_i contains main content, knowledge point labels, academic stage difficulty label g_i and source information.

Different from general Q&A tasks, the task in this paper emphasizes two constraints: first, answers should be generated based on retrieved evidence as much as possible to reduce unsubstantiated expansion; second, the terminology density, explanation depth and example form of answers should match student portraits.

3.2. Construction of Scientific Knowledge Base and Difficulty Labels

The scientific knowledge base is sorted out from primary and secondary school teaching materials covering science, physics, chemistry, biology, geography and information technology. Each knowledge fragment is controlled within 80-180 words, which retains complete conceptual information and facilitates vector retrieval(Wang et al., 2021). To reduce labeling costs, this paper adopts a "rule-based preliminary labeling + manual correction" strategy to generate difficulty labels. The rule-based preliminary labeling determines candidate academic stages based on textbook grades, conceptual keywords and sentence complexity, while manual correction is mainly used to process interdisciplinary concepts and abstract expressions.

The knowledge fragments are divided into three difficulty levels: upper primary school, junior high school and senior high school, encoded as 1, 2 and 3 respectively. For knowledge fragments applicable to multiple academic stages, the lowest intelligible academic stage is selected as the primary label, and the expansion depth is controlled by prompt words in the generation stage. These labels are not treated as simple text-length indicators; instead, they estimate the distance between the knowledge fragment and the learner's current cognitive level, thereby balancing the Zone of Proximal Development and cognitive-load control.

3.3. Difficulty-Aware Retrieval Score

Traditional dense retrieval ranks candidate content based on the cosine similarity of question and document vectors(Karpukhin et al., 2020; Wang et al., 2021). On this basis, this paper adds portrait matching items and difficulty deviation penalty items. The semantic similarity between question q and knowledge fragment d is defined as:

$$\mathrm{Sim}(q,d)=\cos_{f_\theta}(E_q(q),E_d(d)) \tag{1}$$

where E_q and E_d denote the question encoder and document encoder, respectively. Considering the deviation between the student academic stage g_u and document difficulty g_d, the difficulty deviation is defined as:

$$\mathrm{Diff}(d,u)=\frac{|g_d-g_u|}{G} \tag{2}$$

where G represents the maximum difficulty level difference. A larger G indicates a

greater mismatch between the knowledge fragment and the student's current academic stage. If the document labels are consistent with the interest topics, weak knowledge points or current learning units in the student portrait, the portrait matching score Match(d,u) will be higher. The final retrieval score is defined as:

$$Score(q,d,u)=\alpha Sim(q,d)+\beta Match(d,u)-\gamma Diff(d,u) \quad (3)$$

where α, β and γ are weight coefficients controlling semantic relevance, portrait matching and difficulty-deviation penalty, respectively. In the retrieval stage, Top-N candidate fragments are first recalled and then re-ranked according to Score, with Top-k evidence finally selected and input into the generation model. The core advantage of this design is that it does not change the main structure of the retriever, but adds educational constraints in the re-ranking stage, resulting in low implementation cost.

3.4. Evidence-Constrained Answer Generation

Answer generation adopts a structured prompt template, which consists of five parts: task description, student portrait, retrieved evidence, answer requirements and safety constraints. The answer requirements are specified as follows: answers must be generated based on existing evidence first; if evidence is insufficient, "uncertain" should be stated instead of fabricating content; answers for primary school students should avoid abstract terminology and add more life-oriented examples; answers for junior high school students can include basic causal explanations and simple formulas; answers for senior high school students can appropriately introduce mechanism analysis.

The generation model is defined as $y=M_\theta(q,u,C,p)$, where C denotes the set of retrieved evidence and p denotes the prompt template. To enhance the stability of the model's popular science writing style, this paper conducts LoRA fine-tuning with high-quality small-scale popular science Q&A samples. For the pre-trained weight matrix W, a low-rank increment $\Delta W=BA$ is introduced, where $B\in R^{d\times r}$, $A\in R^{r\times k}$, and r is much smaller than d and k. Only matrices A and B are updated during fine-tuning.

$$W'=W+\Delta W=W+BA \quad (4)$$

$$L=L_{ce}+\lambda L_{style} \quad (5)$$

where L_{ce} denotes the standard cross-entropy loss, and L_{style} is used to constrain answer length, terminology density and explanation structure. In practical implementation, L_{style} can be approximated by rule-based indicators or constructed with weakly supervised labels based on human preference samples.

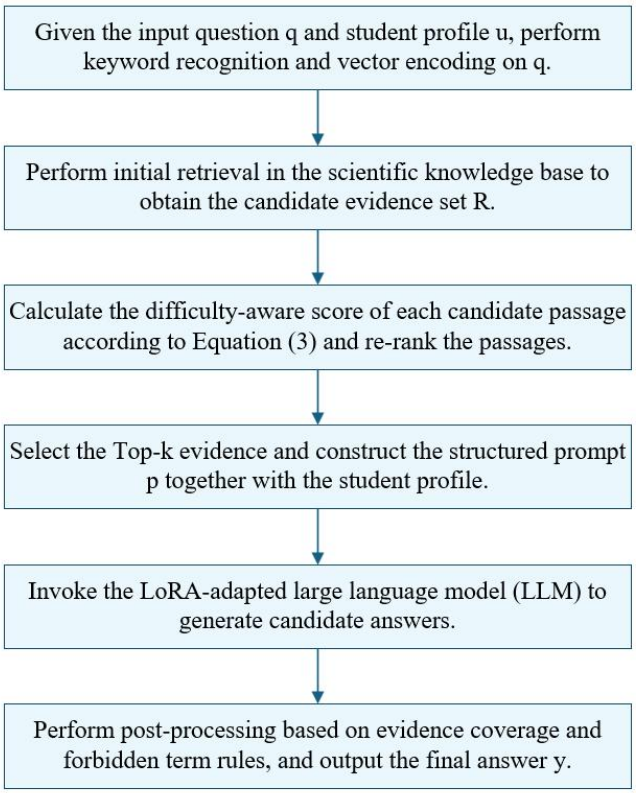
3.5. Algorithm Flow

The algorithm formalizes the end-to-end pipeline of the proposed method. It standardizes key steps from retrieval to generation, ensuring reproducibility.

As illustrated in Figure 2, the procedure sequentially executes difficulty-aware retrieval, prompt construction, LoRA adaptation and answer generation, embedding learner and difficulty constraints throughout.

Figure 2

Flowchart of difficulty-aware retrieval-augmented Q&A generation algorithm



Note. The algorithm pipeline sequentially performs difficulty-aware retrieval, structured prompt construction, LoRA-adapted generation, and post-processing output.

4. Experiments and Result Analysis

4.1. Dataset and Experimental Settings

To verify the effectiveness of the proposed method, this paper constructs the K12-ScienceQA-CN dataset. The dataset covers five themes including physics, chemistry, biology, geography and information technology, containing 600 scientific questions, 600 standard answers and 1800 knowledge fragments. All questions are divided into three academic stages: upper primary school, junior high school and senior high school. The dataset is split into training set, validation set and test set at a ratio of 7:1:2. In the evaluation stage, each question corresponds to a student portrait containing academic stage, knowledge

foundation and expression preference attributes.

Qwen3-8B-Instruct is adopted as the base generator, which is an instruction-tuned model with approximately 8B parameters. A Chinese sentence embedding model is used for vector retrieval. FAISS is utilized to build indexes for vectorized knowledge base fragments(Johnson et al., 2019). The LoRA rank r is set to 8, the learning rate is 2×10^{-4}, the training epoch is 3, and the Top-k evidence number is 4. The default weights of the difficulty-aware retrieval function are set as $\alpha=0.65$, $\beta=0.20$, $\gamma=0.15$.

Table 1

Statistical Information of K12-ScienceQA-CN Dataset

Category	Number of Questions	Number of Knowledge Fragments	Examples of Core Knowledge Points
Physics	150	450	Force, light, sound, electricity, energy conversion
Chemistry	100	300	Material changes, acid-base properties, dissolution, combustion
Biology	150	450	Cells, plants, human body, ecosystem
Geography	100	300	Weather, earth movement, water cycle, terrain
Information Technology	100	300	Algorithms, network, data, security

Note. The dataset covers five K-12 science subjects, including 600 questions and 1800 knowledge fragments in total. All fragments are labeled with academic-stage difficulty levels for difficulty-aware retrieval matching.

4.2. Baseline Methods and Evaluation Metrics

Six baseline methods are set for comparison: (1) Base LLM: direct answer generation by Qwen3-8B-Instruct without an external knowledge base; (2) BM25-RAG(Robertson & Zaragoza, 2009) : retrieval-augmented generation based on keyword matching; (3) Dense-RAG(Karpukhin et al., 2020) : retrieval-augmented generation based on vector similarity; (4) LoRA-LLM(Hu et al., 2022) : pure LoRA fine-tuning for popular science Q&A without retrieval evidence; (5) Profile-RAG: conventional RAG integrated with student portraits in prompt templates without difficulty matching in the retrieval stage; (6) Cross-Encoder-RAG: after Dense-RAG recalls Top-N candidate fragments, a cross-encoder-based neural re-ranker scores each question-evidence pair(Nogueira & Cho, 2019), which is used to evaluate the effectiveness of the proposed difficulty-aware re-ranking mechanism against a stronger neural re-ranking baseline.

The evaluation metrics include scientific accuracy (Acc), age appropriateness score (AgeFit), completeness score (Comp) and hallucination rate (Hallu). Acc is judged by evaluators based on standard scientific answers; AgeFit and Comp are scored manually on a 1-5 scale; Hallu refers to the proportion of answers with obvious fabrication, unsupported content or scientific errors. Each sample is scored independently by three evaluators, and the average value is taken as the final result.

To strengthen the reliability of the human evaluation protocol, inter-rater agreement is further reported. Fleiss' Kappa is used for categorical judgments such as Acc and Hallu(Fleiss, 1971), while the two-way random-effects intraclass correlation coefficient ICC(2,k) is used for 1-5 scale scores such as AgeFit and Comp(Shrout & Fleiss, 1979). The agreement among the three evaluators is κ_{Acc}=0.78, κ_{Hallu}=0.74, ICC_{AgeFit}=0.83 and ICC_{Comp}=0.81, indicating good consistency of the human annotations.

As shown in Table 2, the pure large language model achieves relatively low scientific accuracy and completeness with a high hallucination rate of 15.6%. The introduction of BM25(Robertson & Zaragoza, 2009) or dense retrieval(Karpukhin et al., 2020) significantly reduces hallucinations by providing external evidence. LoRA-LLM(Hu et al., 2022) improves age appropriateness, indicating that small-scale popular science fine-tuning helps the model learn educational expression styles, but factual reliability remains limited without retrieval evidence. Profile-RAG further improves age appropriateness by integrating student portraits. Cross-Encoder-RAG strengthens semantic matching between questions and evidence, achieving an Acc of 91.3%, but it mainly optimizes relevance and does not explicitly model academic-stage difficulty or learner portraits. In contrast, the proposed method jointly considers semantic relevance, portrait matching and difficulty deviation, achieving the best overall performance with an Acc of 92.4%, an AgeFit of 4.36 and a Hallu rate of 5.9%.

Table 2

Quantitative Comparison of Different Methods on Popular Science Q&A Tasks

Method	Acc/%	AgeFit	Comp	Hallu/%
Base LLM	82.3	3.54	3.68	15.6
BM25-RAG	86.7	3.71	3.92	10.8
Dense-RAG	89.1	3.85	4.08	8.4
CrossEnc-RAG	91.3	4.12	4.23	6.5
LoRA-LLM	87.5	4.05	3.97	10.1
Profile-RAG	90.4	4.06	4.14	7.3

Proposed Method	92.4	4.36	4.29	5.9

Note. Acc = scientific accuracy; AgeFit = age appropriateness score; Comp = completeness score; Hallu = hallucination rate. All metrics are evaluated by three independent raters.

4.3. Ablation Experiments

Table 3 shows that removing the difficulty matching item causes a slight drop in scientific accuracy but a significant decrease in age appropriateness (from 4.36 to 3.74), which verifies that the difficulty-aware mechanism mainly optimizes the educational adaptability of answers. Removing retrieval augmentation leads to a sharp rise in the hallucination rate, proving that external evidence is the core guarantee of scientific accuracy. Removing LoRA adaptation reduces the stability of answer style and structural standardization, especially for primary school-oriented questions.

Table 3

Results of Ablation Experiments

Method	Acc/%	AgeFit	Comp	Hallu/%
Full Method	92.4	4.36	4.29	5.9
w/o Difficulty Matching	90.8	3.74	4.21	6.7
w/o Portrait Matching	91.2	3.81	4.22	6.4
w/o Retrieval Augmentation	84.9	4.02	3.83	13.1
w/o LoRA Adaptation	90.7	4.01	4.18	7.0

Note. Each row removes one core module from the full proposed method to verify its independent contribution.

4.4. Parameter Sensitivity Analysis

Table 4 indicates that an excessively high α makes the system approximate conventional semantic retrieval, limiting the improvement of age appropriateness. An excessively high γ causes the model to over-prefer simple knowledge materials, excluding key valid evidence and reducing scientific accuracy.

Notably, when $\gamma=0.25$, AgeFit further increases to 4.45, whereas Acc drops to 90.9%. This unexpected finding shows that "easier" evidence is not always pedagogically better; popular science Q&A systems must balance understandability and scientific completeness rather than simply simplifying the text. Considering the comprehensive performance of accuracy and age appropriateness, $\alpha=0.65$, $\beta=0.20$ and $\gamma=0.15$ are adopted as the default

configuration.

Table 4

Sensitivity Analysis of Retrieval Score Weights

α	β	γ	Acc/%	AgeFit
0.80	0.10	0.10	91.8	4.05
0.65	0.20	0.15	92.4	4.36
0.55	0.30	0.15	91.7	4.40
0.55	0.20	0.25	90.9	4.45

Note. α = weight of semantic relevance; β = weight of portrait matching; γ = weight of difficulty deviation penalty.

4.5. Inter-Rater Reliability and Error Analysis

Beyond the overall metrics, we further analyze the question types on which the model still underperforms. Errors mainly occur in three types of questions: cross-disciplinary and multi-step causal reasoning questions, such as explanations involving both physical and chemical changes; questions requiring precise conditional constraints or numerical relationships, such as why the temperature of boiling water remains nearly unchanged under standard atmospheric pressure; and questions whose necessary evidence is distributed across multiple fragments, where the model may rely on only one fragment and ignore key conditions.

A typical incorrect case is as follows. For the question “Why does the temperature of water basically stop rising after it boils, even if heating continues?” with an upper-primary-school learner portrait, one generated answer was: “Because the water is already very hot, the temperature will still rise slowly if heating continues, but the increase is not obvious.” Although the wording is simple, the answer ignores latent heat of vaporization and may mislead students. A more appropriate explanation should state that, under standard atmospheric pressure, the heat absorbed after boiling is mainly used to convert liquid water into water vapor, so the temperature usually remains near 100°C. The error occurs because retrieval fragments contain keywords such as “heating”, “temperature increase” and “boiling” simultaneously, and the model tends to follow everyday-experience reasoning instead of sufficiently using the key evidence about vaporization.

5. Conclusion

This paper proposes a difficulty-aware retrieval-augmented personalized popular science

Q&A generation method for science education scenarios. Based on the classic RAG framework(Lewis et al., 2020; Gao et al., 2023), the proposed method integrates student portrait matching and knowledge difficulty deviation penalty, enabling the retrieval process to optimize both semantic relevance and educational applicability. Meanwhile, structured prompts and LoRA parameter-efficient fine-tuning(Hu et al., 2022) are adopted to enhance the model's popular science expression capability. Experimental results demonstrate that the proposed method outperforms baseline methods in scientific accuracy, age appropriateness and hallucination control, which is suitable for low-computational-cost academic research and educational prototype system development.

The main findings suggest that the advantage of the proposed method does not simply come from stronger language generation, but from the joint retrieval constraint of relevance, difficulty and learner portrait. The difficulty-matching term substantially improves AgeFit, retrieval augmentation reduces Hallu, and LoRA adaptation enhances the stability of answer structure and writing style. These results indicate that educational RAG systems should not only retrieve semantically relevant evidence, but also explicitly incorporate learners' cognitive levels into evidence selection and answer generation.

This study also has limitations. First, the K12-ScienceQA-CN dataset is still limited in scale, and academic-stage labels are divided into only three levels, which cannot fully characterize individual differences among students. Second, experiments are mainly conducted in Chinese science Q&A scenarios, so the generalization ability across textbook versions, regional curricula and multilingual settings requires further validation. Third, questions involving multi-step reasoning, implicit conditions and numerical relationships may still suffer from insufficient evidence utilization or incomplete explanation. The parameter sensitivity analysis also reveals an unexpected trade-off: excessive preference for low-difficulty evidence can improve surface-level age appropriateness but weaken scientific completeness.

In terms of practical implications and social significance, the proposed method can provide a low-cost personalized Q&A solution for K12 popular science platforms, intelligent tutoring systems and teacher preparation tools. It may help students at different academic stages obtain explanations that better match their cognitive levels, while evidence constraints and hallucination control improve the trustworthiness of educational answers. Nevertheless, the system should be regarded as an assistive tool for teachers and learners rather than a replacement for teachers' professional judgment about students' understanding, classroom context and value guidance.

Future research can be expanded from three aspects: first, expanding the scale of the

knowledge base and introducing textbook version information to improve retrieval evidence coverage; second, constructing fine-grained cognitive learner models to realize multi-dimensional personalized adaptation beyond single academic stage differentiation; third, introducing automatic factual verification and teacher feedback closed-loop mechanisms to further improve the reliability of the system in real classroom scenarios.

Acknowledgments

The authors have no additional acknowledgments to declare.

Author contributions

Conceptualization, Yuewen Cao; methodology, Yuewen Cao; software, Yuewen Cao and Xinyang Li; validation, Yuewen Cao, Xinyang Li, Chunyu Lei and Lin Zhang; formal analysis, Yuewen Cao; investigation, Yuewen Cao and Xinyang Li; resources, Chunyu Lei and Lin Zhang; data curation, Yuewen Cao and Xinyang Li; writing—original draft preparation, Yuewen Cao; writing—review and editing, Yuewen Cao, Xinyang Li, Chunyu Lei and Lin Zhang; visualization, Yuewen Cao; supervision, Yuewen Cao and Lin Zhang; project administration, Yuewen Cao; funding acquisition, Yuewen Cao and Lin Zhang. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the 2026 Undergraduate Innovation and Entrepreneurship Training Program of Beijing Technology and Business University, grant number S202610011034.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Conflict of interest

The authors declare that there is no conflict of interest.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

All authors have given their consent.

WISDOM ACADEMIC https://doi.org/10.63808/css.v2i3.426 Volume 2 , Issue 3

Additional information

Received: 25 May 2026

Accepted: 18 June 2026

Published online: 16 August 2026

Publisher's Note

References

[1] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., … Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901).

[2] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 24824–24837).

[3] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459–9474).

[4] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6769–6781). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.550

[5] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

[6] Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (pp. 4582–4597). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.353

[7] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 2790–2799). PMLR. http://proceedings.mlr.press/v97/houlsby19a.html

[8] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

[9] Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval, 3*(4), 333–389. https://doi.org/10.1561/1500000019

[10] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2312.10997

[11] Mallen, A., Asai, A., Zhong, V., Pérez, E., & Hajishirzi, H. (2023). When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (pp. 9802–9822). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.546

[12] 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., Stadler, M., Weller, J., Kuhn, J., & Winne, P. H. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

[13] Zhai, X. (2022). ChatGPT user experience: Implications for education. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4312418

[14] Wang, S., Scells, H., Zuccon, G., & Azzopardi, L. (2021). Neural ranking models for document retrieval: A survey. *Information Retrieval Journal, 24*, 349–388. https://doi.org/10.1007/s10791-021-09396-7

[15] OpenAI. (2023). GPT-4 technical report [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2303.08774

[16] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *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. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

[17] Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data, 7*(3), 535–547. https://doi.org/10.1109/TBDATA.2019.2956512

[18] Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

[19] Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

[20] Nogueira, R., & Cho, K. (2019). Passage re-ranking with BERT [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1901.04085

[21] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382. https://doi.org/10.1037/h0031619

[22] Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420