On the Robustness of RAG Systems in Educational Question Answering under Knowledge Discrepancies

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) systems have demonstrated remarkable potential as question answering systems in the K-12 Education domain, where knowledge is typically queried within the restricted scope of authoritative textbooks. However, the discrepancy between textbooks and the parametric knowledge in Large Language Models (LLMs) could undermine the effectiveness of RAG systems. To systematically investigate the robustness of RAG systems under such knowledge discrepancies, we present EDUKDQA, a question answering dataset that simulates knowledge discrepancies in real applications by applying hypothetical knowledge updates in answers and source documents. EDUKDQA includes 3,005 questions covering five subjects, under a comprehensive question typology from the perspective of context utilization and knowledge integration. We conducted extensive experiments on retrieval and question answering performance. We find that most RAG systems suffer from a substantial performance drop in question answering with knowledge discrepancies, while questions that require integration of contextual knowledge and parametric knowledge pose a challenge to LLMs. All resources will be released to foster further research.

1 Introduction

In K-12 education, Question Answering (QA) systems serve as an important resource for learning assistance, where answers are precisely provided within a restricted knowledge scope from authoritative sources (i.e., textbooks) (Raamadhurai et al., 2019; Soares et al., 2021). Meanwhile, benefiting from the emergent abilities (Wei et al., 2022) of Large Language Models (LLMs) and advanced information retrieval (IR) methods, Retrieval-Augmented Generation (RAG) systems have achieved remarkable performance in various knowledge-intensive tasks in natural language processing (Lewis et al., 2020; Jiang et al., 2023; Gao et al., 2024), demonstrating their great potential as QA systems in K-12 education (Gan et al., 2023; Kasneci et al., 2023; Yan et al., 2024).

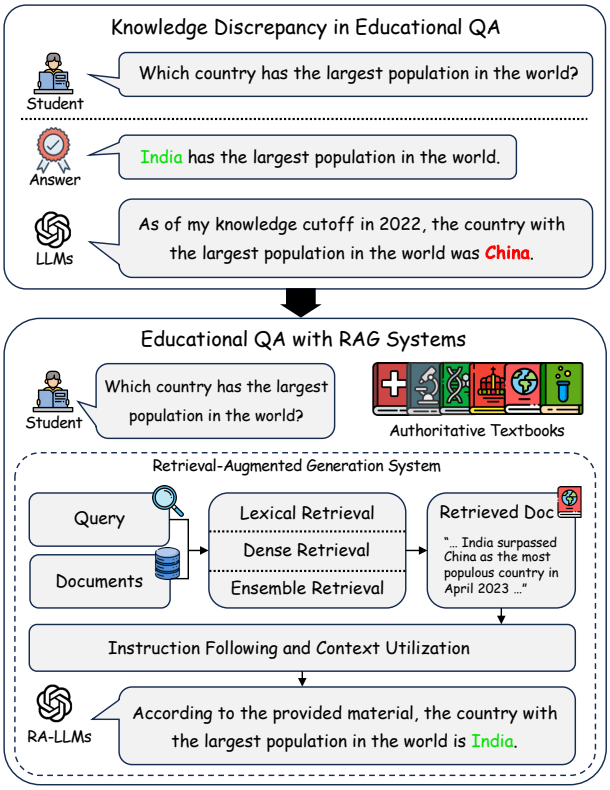
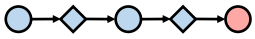
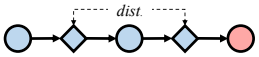
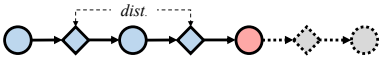


Figure 1: An illustration of knowledge discrepancy in educational QA and the application of RAG systems.

In K-12 educational QA, one of the primary concerns is ensuring that the knowledge conveyed in the answer is consistent with the officially designated textbooks (Extance, 2023). However, there are notable discrepancies between the knowledge in textbooks and the internal knowledge of LLMs, due to the evolving nature of facts (Arbesman, 2012), updates in pedagogical approaches (Provenzo et al., 2011), as well as regional and cultural variations in content (Patel, 2015). It remains unclear whether RAG systems can robustly incorporate knowledge from authoritative sources and generate consistent answers under such knowledge discrepancies (sce-

Question Type	Reasoning Pattern and Example Question	Hypothetical Knowledge Update
Simple Direct		NV goggles - detect - ~~Infrared~~ light ***Ultraviolet***
	What type of light is *detected* by **night vision goggles**?	Original: Infrared Light Updated: Ultraviolet Light
Multi-hop Direct		de Broglie eq. - developed by - ~~Louis~~ de Broglie ***Maurice***
	Which scientist *developed* an **equation** that can *calculate* the **wavelength of a particle**?	Original: Louis de Broglie Updated: Maurice de Broglie
Multi-hop Distant		~~Na^+/K^+~~ Pump - creates - EC Gradient ***Ca^{2+}***
	Which pump *creates* an **electrochemical gradient** that *enables* **secondary active transport** to occur?	Original: Sodium-potassium Pump Updated: Calcium Pump
Multi-hop Implicit		~~Polonium~~ - found in - Uranium ores ***Thorium***
	Who *discovered* the **radioactive element** that is commonly *found* in **uranium ores**?	Original: Marie Curie Updated: Jöns Jacob Berzelius
Distant Implicit		~~Mitochondrion~~ - conducts - Cellular respiration ***Golgi apparatus***
	Who *discovered* the **organelle** that is *responsible* for the **biological process** that *produces* **ATP**?	Original: Albert von Kölliker Updated: Camillo Golgi

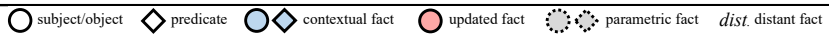


Table 1: Five question types in EDUKDQA with their reasoning patterns, example questions, and hypothetical knowledge update illustrated in factual triplets and answers. In example questions, subjects and objects are marked in **bold**, while predicates are marked in *italic*.

nario illustrated in Figure 1).

To fill this gap, we aim to systematically assess the robustness of RAG systems in performing question answering in K-12 education when encountering knowledge discrepancies. We present **EDUKDQA** (**Edu**cational **K**nowledge **D**iscrepancy **Q**uestion **A**nswering), a new dataset containing 3,005 multiple-choice questions covering the subjects of Physics, Chemistry, Biology, Geography, and History from the middle-school curriculum. To simulate the knowledge discrepancy between LLMs and textbooks, we conduct a *hypothetical knowledge update*, in which we modify the original factual knowledge from the textbooks into plausible alternatives while maintaining coherent and consistent context. Moreover, we tailored a comprehensive question typology to stress-test the *context utilization* and *knowledge integration* abilities of LLMs under such scenarios.

We conducted extensive experiments with various retrieval methods and LLMs. We find that most RAG systems still suffer from a considerable performance drop when facing knowledge discrepancies. Notably, while most LLMs can incorporate distant contextual facts well, they struggle to integrate their parametric knowledge with contextual knowledge effectively. In terms of retrieval, traditional lexical-based methods show advantages due to specificity of the academic terms, and their performance may be further enhanced through an ensemble reranking mechanism. To encourage further research, we will make our benchmark dataset and the associated code publicly available.

2 The EDUKDQA Dataset

In this section, we introduce the methodology of hypothetical knowledge update and the design of our question typology. The curation pipeline and detailed statistics of the EDUKDQA dataset are provided in Appendix B and E, respectively.

2.1 Hypothetical Knowledge Update

One of the core objectives of our dataset is to simulate the knowledge discrepancy between LLMs and authoritative textbooks when performing educational QA. However, such discrepancies are often fuzzy and highly sparse in real-world data, making it infeasible to collect and organize. Consequently, we designed the methodology of hypothetical knowledge update, performing it on high-quality open-source textbooks. The general procedures are as follows: 1) Curate factual questions from textbook paragraphs following our designed question typology. 2) Select a plausible but factually incorrect answer as the updated ground-truth answer. 3) Replace all occurrences of the original answer in the paragraph with the updated answer and adjusted other relevant statements in the

context to ensure that the updated paragraph is coherent and consistent. This process is further guaranteed through extensive human curation and verification. Examples of hypothetical knowledge updates in our dataset are provided in Appendix D.

2.2 Question Typology

We identify two potential challenges for LLMs when performing QA with knowledge discrepancy: 1) *Context Utilization*: Whether LLMs can identify and utilize the corresponding facts from the context; and 2) *Knowledge Integration*: Whether LLMs can incorporate their own parametric knowledge with the contextual ones in question answering. To this end, we designed our question typology, as illustrated in Table 1, to investigate such abilities. Based on the two basic question types, *Simple Direct* and *Multi-hop Direct*, we developed the *Multi-hop Distant* type to evaluate the context utilization ability for distant facts from the passage, and the *Multi-hop Implicit*[1] type to evaluate the knowledge integration ability that combines their own factual knowledge with retrieved ones. Moreover, the *Distant Implicit* type poses a greater challenge by combining both features. To ensure our evaluation of knowledge integration ability is independent of knowledge coverage, we restrict the facts requiring LLMs' own knowledge to be high-frequency only (Sun et al., 2024).

3 Experiments and Analyses

Typically, RAG systems first conduct document retrieval based on given queries, then perform question answering with LLMs based on the retrieved information loaded in the context. In this section, we comprehensively evaluate and analyze the performance of retrieval methods and LLMs on the EDUKDQA dataset. For details of all tested methods and models, please refer to Appendix C.

3.1 Retrieval Performance

Experimental results of the retrieval methods are presented in Table 2. Traditional lexical retrieval methods, such as BM25, demonstrated strong performance on our dataset, while dense retrieval methods, such as Mistral-embed and Ada-002, achieved comparable performance. Since our dataset focuses on the K-12 education domain, lexical retrieval effectively captures domain-specific keywords in the queries and identifies the corresponding documents. This characteristic underscores the need for fine-tuning dense retrieval models on our dataset. To this end, we fine-tuned both Contriever and Contriever-msmarco on our documents, resulting in significant improvements and highlighting the importance of corpus-specific fine-tuning in educational document retrieval.

Retrieval Methods	Category	R@1	R@5
TF-IDF (Spärck Jones, 1972)	Lexical	65.82	88.72
BM25 (Robertson et al., 1994)	Lexical	82.73	95.27
SPLADE (Formal et al., 2021)	Lexical/Dense	78.04	90.12
Contriever (Izacard et al., 2022)	Dense	53.18	81.80
Con.-msmarco (Izacard et al., 2022)	Dense	76.17	93.54
Mistral-embed (Mistral, 2023a)	Dense	78.74	95.31
Ada-002 (OpenAI, 2022)	Dense	79.23	95.44
Query Rewrite (Ma et al., 2023)	Pre-Retrieval	78.87	94.21
Hybrid Rerank (BM25 + Ada-002)	Ensemble	<u>84.43</u>	<u>96.04</u>
Contriever (fine-tuned)	Dense+FT	<u>84.19</u>	<u>98.96</u>
Con.-msmarco (fine-tuned)	Dense+FT	**87.95**	**99.50**

Table 2: Document retrieval performances (in recall @1/5 %) of retrieval methods from different categories in the EDUKDQA dataset. The highest scores are marked as **bold**, while the 2nd and 3rd-best scores are <u>underlined</u>. The retrieval granularity is set to paragraph.

Moreover, ensemble methods have demonstrated their effectiveness in various information retrieval tasks (Thakur et al., 2021). We implemented a hybrid approach that retrieves the top k[2] documents with Ada-002, followed by re-ranking with BM25. This ensemble method marginally improved the retrieval performance in both metrics. Conversely, query rewriting[3] (Ma et al., 2023) implemented upon BM25 did not yield performance gains. Detailed results across subjects and question types are provided in Appendix F.

3.2 Question Answering Performance

Experimental results of LLMs in question answering are presented in Table 3. Most LLMs exhibited comparable performance for question types *Simple Direct*, *Multi-hop Direct*, and *Multi-hop Distant*. This suggests that **both multi-hop reasoning and distant context utilization do not pose significant challenges for modern LLMs**. However, for *Multi-hop Implicit* questions requiring the integration of contextual and internal knowledge, a substantial performance disparity emerges between smaller open-source LLMs and advanced models. This disparity is further amplified for *Distant Implicit* questions. The most capable model, o1-preview, attains over 80% accuracy on *Implicit* questions, whereas Mistral-7b's performance

[1]The word 'implicit' indicates that the questions indirectly query the updated facts by embedding them within the middle of the multi-hop reasoning chain.

[2]We use $k = 6$ as the optimal hyperparameter setting.

[3]We use `Mistral-small-2409` as the LLM query rewriter.

Large Language Models	Question Typology					Average
	Simple Direct	*Multi-hop Direct*	*Multi-hop Distant*	*Multi-hop Implicit*	*Distant Implicit*	
Mistral-7b (Mistral, 2023b)	77.70	69.31	72.74	45.32	33.98	61.26
Mixtral-8x22b (Mistral, 2023c)	84.10	84.58	87.15	73.86	65.37	79.67
Mistral-small-2409 (Mistral, 2024a)	87.87	88.70	89.69	72.18	60.81	80.77
Mistral-large-2407 (Mistral, 2024b)	83.93	83.51	87.29	82.25	70.57	81.66
Gemini-1.5-flash (Google, 2024a)	80.98	82.44	87.57	76.02	63.25	78.54
Gemini-1.5-pro (Google, 2024b)	86.56	87.63	88.42	76.26	63.58	81.10
Llama3-8b (Meta, 2024)	90.33	88.85	89.69	63.55	49.43	77.77
Llama3-70b (Meta, 2024)	96.72	96.49	96.89	79.14	63.25	87.42
GPT-3.5-turbo (OpenAI, 2022)	92.62	90.53	91.24	71.22	57.72	81.73
GPT-4 (OpenAI, 2023a)	89.51	89.47	90.68	78.66	70.89	84.46
GPT-4-turbo (OpenAI, 2023b)	95.74	96.18	96.19	81.06	71.71	88.99
GPT-4o (OpenAI, 2024a)	91.97	94.81	93.64	81.29	70.41	87.09
Claude-3-sonnet (Anthropic, 2024a)	94.59	92.82	93.64	77.70	60.65	84.69
Claude-3.5-sonnet (Anthropic, 2024b)	**97.54**	96.49	95.62	83.69	73.82	90.08
o1-mini (OpenAI, 2024b)	95.90	95.73	97.03	85.85	75.45	90.55
o1-preview (OpenAI, 2024b)	95.08	**97.71**	**97.46**	**86.33**	**78.86**	**91.68**

Table 3: Question answering performances (in accuracy %) of LLMs in the EDUKDQA benchmark with corresponding documents provided. The highest scores are marked as **bold**, while the 2nd and 3rd-best scores are underlined. All LLMs are tested under zero-shot settings, with a *Locate-and-Answer* prompting approach that facilitates active information acquisition from contextual documents, with details in Appendix G.

RAG System	Hypothetical Knowledge Update		Drop
	Before	After	
Llama3-8b + Ada-002	87.49	62.60	24.89
Llama3-8b + Rerank	88.49	66.02	22.47
GPT-4o + Ada-002	96.57	69.65	26.92
GPT-4o + Rerank	97.10	73.71	23.39

Table 4: Performance drop of RAG systems with hypothetical knowledge updates in our benchmark.

falls below 40%. These findings indicate that **knowledge integration is an emergent capability presenting greater difficulties for LLMs under knowledge discrepancies**, particularly when coupled with the need for distant context utilization.

3.3 Overall Performance

How do knowledge discrepancies affect the performance of RAG systems in educational application? We selected two representative LLMs: Llama3-8b from open-source models and GPT-4o for proprietary models, combined with two high-performing retrieval methods: Ada-002 and hybrid rerank. The resulting RAG systems were tested on questions *before* and *after* hypothetical knowledge updates, with results presented in Table 4. We observed an accuracy drop of 22-27%, indicating a substantial performance degradation in modern RAG systems when faced with knowledge discrepancies.

4 Related Work

Retrieval-Augmented Generation Following the categorization by Gao et al. (2024), the RAG methods in our experiment are from *Naive RAG* (lexical, dense) and *Advanced RAG* (rerank, rewrite). Recently, *Modular RAG* has emerged to enhance the adaptability and versatility of RAG systems (Shao et al., 2023; Asai et al., 2023).

Educational Question Answering Prior to the emergence of RAG systems utilizing LLMs, various educational QA systems were developed to provide pedagogically appropriate responses to student inquiries (Abdi et al., 2018; Agarwal et al., 2019). While recent literature explores LLM applications in QA and learning assistance roles (Nye et al., 2023; Kuo et al., 2023; Wang et al., 2024).

Knowledge Discrepancy in LLMs Mitigating knowledge discrepancies or conflicts in applications is a fundamental challenge in LLM research (Xu et al., 2024). Researchers have proposed tuning-based (Li et al., 2022) and prompting-based (Zhou et al., 2023) methods to enhance LLMs' robustness under such conflicts. However, these conflicts in educational applications and RAG systems remain relatively underexplored.

5 Conclusion

This paper systematically evaluates the robustness of RAG systems in K-12 educational question answering under knowledge discrepancies using a comprehensive dataset. Experimental findings reveal substantial performance degradation in RAG systems when faced with knowledge discrepancies, which is primarily attributed to deficiencies in incorporating contextual and parametric knowledge in question answering—an emergent and challenging ability for modern large language models.

Limitations

We discuss three main limitations of our work.

First, EDUKDQA employs the approach of hypothetical knowledge updates to effectively simulate real-world knowledge discrepancies for two primary reasons: (1) Real-world knowledge conflicts are often sparse, noisy, and difficult to systematically collect or organize into a cohesive dataset suitable for large-scale evaluation. Hypothetical updates provide a scalable and controlled alternative, allowing us to bypass these limitations. (2) By leveraging a systematic annotation and curation pipeline, we can generate questions with diverse and well-defined reasoning patterns that align with our typology, enabling more robust evaluation of complex question-answering tasks. For future research centered on real-world knowledge discrepancies, we recommend an alternative methodology that incorporates temporal attributes (Chen et al., 2021; Zhang and Choi, 2024). This approach focuses on identifying outdated facts and capturing time-sensitive data (e.g., economic trends, annual events, or societal changes) to construct datasets that reflect real-world knowledge updates. While promising, this method is constrained by the limited overlap between time-sensitive data—often numerical or attribute-specific—and the broader contextual needs of educational question-answering tasks, which may reduce the comprehensiveness of the resulting datasets.

Next, regarding document retrieval, some recent hierarchical retrieval paradigms, such as GraphRAG (Edge et al., 2024) and HippoRAG (Gutiérrez et al., 2024), are not included in our experiments due to their implementation complexity. However, we believe that such structured paradigms could effectively enhance retrieval performance in our scenario, as educational documents are well-structured and contain high-quality factual knowledge.

Finally, this paper primarily evaluates the robustness of RAG systems in the proposed scenario, with experiments conducted using various retrieval methods and large language models. Potential improvements could be achieved through the design of tailored reasoning frameworks via prompting, in-context learning or alignment in LLMs, which we leave for future research.

Ethics Statement

In constructing the EDUKDQA dataset, we collected text from open-access textbooks. Detailed sources and licenses are provided in Appendix A. The human curation and verification in our annotation pipeline were carried out by a group of postgraduate students with extensive experience in NLP research. We ensured that the updated knowledge is free from harmful or toxic content. It is important to note that our dataset is designed solely to evaluate the robustness of Retrieval-Augmented Generation systems under scenarios with knowledge discrepancies and is not suitable for assessing the factual accuracy of QA systems.

References

Asad Abdi, Norisma Idris, and Zahrah Ahmad. 2018. Qapd: an ontology-based question answering system in the physics domain. *Soft Computing*, 22:213–230.

A. Agarwal, N. Sachdeva, R. K. Yadav, V. Udandarao, V. Mittal, A. Gupta, and A. Mathur. 2019. Eduqa: Educational domain question answering system using conceptual network mapping. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8137–8141.

Anthropic. 2024a. Blog: Introducing the next generation of claude.

Anthropic. 2024b. Introducing claude 3.5 sonnet.

S. Arbesman. 2012. *The Half-life of Facts: Why Everything We Know Has an Expiration Date*. Human body. Current.

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

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization.

Andy Extance. 2023. ChatGPT has entered the classroom: how LLMs could transform education. *Nature*, 623:474–477.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade v2: Sparse lexical and expansion model for information retrieval.

Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Jerry Chun-Wei Lin. 2023. Large language models in education: Vision and opportunities.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey.

Google. 2024a. Gemini flash - google deepmind.

Google. 2024b. Gemini pro - google deepmind.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning.

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.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Bor-Chen Kuo, Frederic T. Y. Chang, and Zong-En Bai. 2023. Leveraging llms for adaptive testing and learning in taiwan adaptive learning platform (talp). In *Empowering Education with LLMs – the Next-Gen Interface and Content Generation*, volume 3487, pages 1–14, Tokyo, Japan. CEUR Workshop Proceedings, CEUR-WS.org.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2022. Large language models with controllable working memory. *arXiv preprint arXiv:2211.05110*.

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

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date.

Mistral. 2023a. Embeddings | mistral ai large language models.

Mistral. 2023b. Mistral 7b | frontier ai in your hands.

Mistral. 2023c. Mixtral of experts | frontier ai in your hands.

Mistral. 2024a. Ai in abundance | mistral ai | frontier ai in your hands.

Mistral. 2024b. Large enough | mistral ai | frontier ai in your hands.

Benjamin D. Nye, Dillon Mee, and Mark G. Core. 2023. Generative large language models for dialog-based tutoring: An early consideration of opportunities and concerns. In *Empowering Education with LLMs – the Next-Gen Interface and Content Generation*, volume 3487, pages 1–15, Tokyo, Japan. CEUR Workshop Proceedings, CEUR-WS.org.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

OpenAI. 2022. New and improved embedding model.

OpenAI. 2023a. Gpt-4 technical report.

OpenAI. 2023b. New models and developer products announced at devday.

OpenAI. 2024a. Hello gpt-4o.

OpenAI. 2024b. Introducing openai o1.

Leigh Patel. 2015. *Decolonizing Educational Research: From Ownership to Answerability*, 1 edition. Routledge, New York.

E.F. Provenzo, A.N. Shaver, and M. Bello. 2011. *The Textbook as Discourse: Sociocultural Dimensions of American Schoolbooks*. Taylor & Francis.

Srikrishna Raamadhurai, Ryan Baker, and Vikraman Poduval. 2019. Curio SmartChat : A system for natural language question answering for self-paced k-12 learning. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 336–342, Florence, Italy. Association for Computational Linguistics.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*. NIST.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy.

Teotino Gomes Soares, Azhari Azhari, Nur Rokhman, and E Wonarko. 2021. Education question answering systems: a survey. In *Proceedings of The International MultiConference of Engineers and Computer Scientists*.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2024. Head-to-tail: How knowledgeable are large language models (LLMs)? A.K.A. will LLMs replace knowledge graphs? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 311–325, Mexico City, Mexico. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. 2024. Large language models for education: A survey and outlook.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey.

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112.

Michael J. Q. Zhang and Eunsol Choi. 2024. Mitigating temporal misalignment by discarding outdated facts.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models.

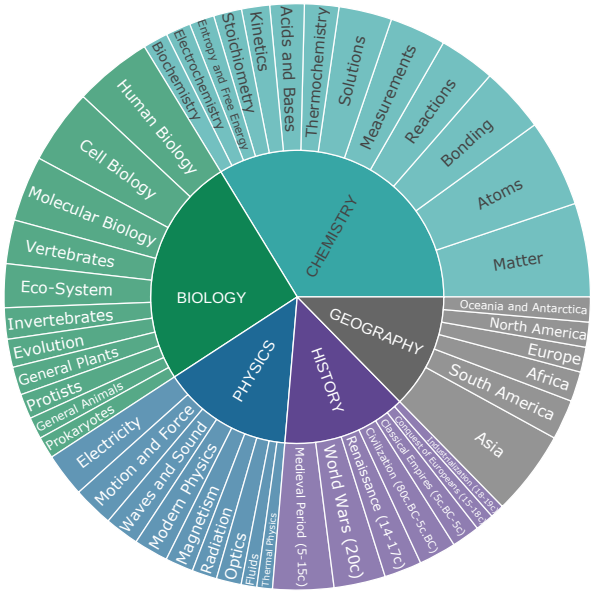


Figure 2: Distribution of five subjects and their corresponding topics included in the EDUKDQA dataset.

A Source Textbooks

The documents in the EDUKDQA dataset is organized based on the following public textbooks:

Physics *Physics* by Openstax (CC-BY-4.0[4]) https://openstax.org/details/books/physics

Chemistry *Chemistry* by Openstax (CC-BY-4.0[1]) https://openstax.org/details/books/chemistry-2e

Biology *Biology* by Openstax (CC-BY-4.0[1]) https://openstax.org/details/books/biology-2e

History *World History* by OER Commons (CC-BY-NC-4.0[5]) https://oercommons.org/courses/world-history-2

Geography *World Regional Geography* by Sailor Academy (CC-BY-3.0[6]) https://learn.saylor.org/course/view.php?id=722

Meanwhile, the detailed topics included in each subjects are illustrated in Figure 2.

[4]https://creativecommons.org/licenses/by/4.0/deed.en
[5]https://creativecommons.org/licenses/by-nc/4.0/deed.en
[6]https://creativecommons.org/licenses/by/3.0/

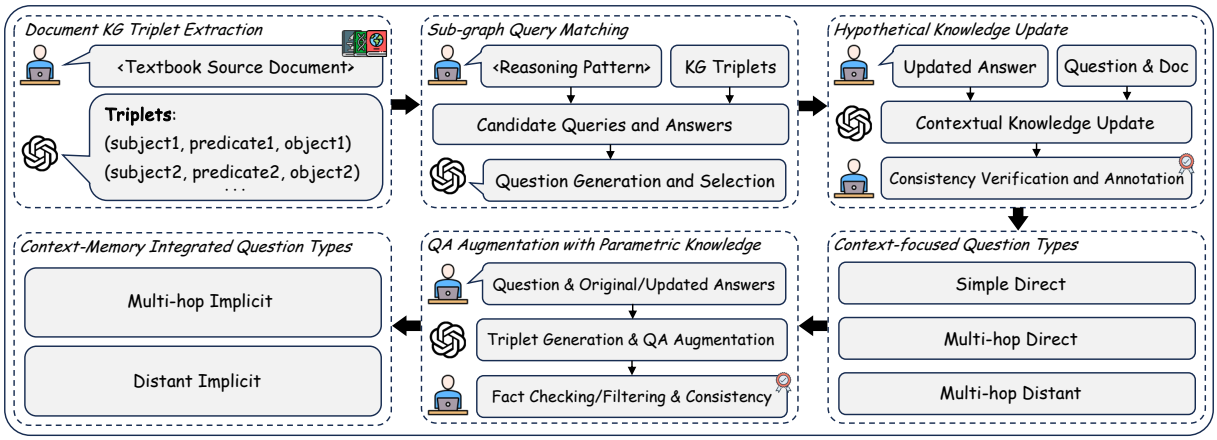


Figure 3: An overview of the data curation pipeline of the EDUKDQA dataset.

B Curation Pipeline

The curation pipeline of our dataset is illustrated in Figure 3. We first perform triplet extraction on the textbook documents and generate a document-level knowledge graph (KG). Next, we perform sub-graph matching based on fixed reasoning patterns to sample candidate queries, and selectively transform them into natural language questions. Then, hypothetical knowledge update is executed and verified to guarantee consistency between the updated answer and the document.

Context-focused question types, including *Simple Direct*, *Multi-hop Direct*, and *Multi-hop Distant*, are acquired through this process. For *Multi-hop Distant* questions, we leverage distant facts, defined as connected triplet pairs that are separated in the document's sequential ordering. These questions are only assigned to documents containing no fewer than 200 words. To generate the other two *Implicit* question types that require knowledge integration, we perform extra QA augmentation followed by an expert verification process.

Our data curation process is performed through an integrated framework involving both human annotators and LLMs. For LLM annotation, we adopted `Claude-3.5-Sonnet` for its outstanding instruction-following ability. Following manual verification, 90.5% of these queries were retained or underwent minor refinements to become high-quality questions, yielding an overall success rate of 86.4%. The total API cost for data annotation is approximately 300 USD.

C Model Details

In this section, we briefly introduce all tested retrieval methods and large language models in our benchmark experiment.

In the retrieval stage of our experiments, we employ a diverse range of retrieval methods. For traditional lexical retrieval, we include **TF-IDF** (Spärck Jones, 1972), which vectorizes documents and queries based on term frequency and inverse document frequency, and **BM25** (Robertson et al., 1994), which enhances TF-IDF with document length normalization and probabilistic term weighting. We also incorporate **SPLADE** (Formal et al., 2021), a method that bridges lexical and dense retrieval paradigms. For dense retrieval, we evaluate several methods that encode questions and documents into the same vector space: **Contriever** (Izacard et al., 2022), an unsupervised text encoder; its fine-tuned variant **Contriever-msmarco**, which we further enhanced by applying contrastive learning (Izacard et al., 2022) to fine-tune both models on the EDUKDQA dataset for improved retrieval capability; and two closed-source embedding models, **Mistral-embed** (Mistral, 2023a) and **Ada-002** (OpenAI, 2022). Additionally, we explore a pre-retrieval method, **Query Rewrite** (Ma et al., 2023), which reformulates queries to improve retrieval performance. Finally, we implement a **Hybrid Rerank** approach that combines BM25 and Ada-002, leveraging the strengths of both lexical and dense retrieval methods.

In our evaluation, we employ a diverse set of state-of-the-art large language models to assess their performance across various tasks. The models include: Mistral AI's open-source models, ranging from the compact **Mistral-7b** (Mistral, 2023b) to the more advanced MoE model **Mixtral-8x22b** (Mistral, 2023c), and their latest iterations **Mistral-small-2409** (Mistral, 2024a) and **Mistral-large-2407** (Mistral, 2024b). Google's Gemini models are represented by **Gemini-1.5-flash**

(Google, 2024a) and **Gemini-1.5-pro** (Google, 2024b). Meta's **Llama3** series is included with both 8b and 70b parameter versions (Meta, 2024). We also evaluate OpenAI's models, including **GPT-3.5-turbo** (OpenAI, 2022), **GPT-4** (OpenAI, 2023a), **GPT-4-turbo** (OpenAI, 2023b), and **GPT-4o** (OpenAI, 2024a). Anthropic's LLMs are represented by **Claude-3-sonnet** (Anthropic, 2024a) and **Claude-3.5-sonnet** (Anthropic, 2024b). Lastly, we include OpenAI's latest o1 series, **o1-mini** and **o1-preview** (OpenAI, 2024b), which achieve remarkable performance across various metrics through inference-time scaling. This comprehensive selection allows us to compare a wide range of model architectures and sizes, providing valuable insights into the current state of LLM capabilities.

D Dataset Example

We provide an example question and its corresponding paragraph before and after hypothetical knowledge update in Table 5.

E Dataset Statistics

The statistics for the documents and questions in the EDUKDQA dataset are provided in Table 6 and Table 7, respectively.

F Result Details

The comprehensive evaluation results for retrieval approaches and large language models across various subjects and question types are presented in Table 8 and Table 9, respectively. We observed that the performance of lexical retrieval methods correlates positively with question length, while dense retrieval methods exhibit an inverse relationship. This finding suggests that there is potential for developing more sophisticated ensemble methodologies that could fully leverage the strengths of both approaches.

G Prompting Approach

In our experiments, we adopt the prompting approach of *Locate-and-Answer*, to facilitate active acquisition of information in the context when performing question answering. We first request LLMs to identify and locate the corresponding sentence that include the knowledge for the question from the provided document, and then reason to provide its answer. According to the experimental result in Table 3, this prompting approach can effectively improve the QA performance of LLMs compared to direct answering.

H Calibration-Induced Performance Discrepancies

In our study, retrieved documents serve as the unequivocal reference, mirroring practices in K-12 education. However, the evaluation results reveal counterintuitive calibration patterns in state-of-the-art LLMs, particularly regarding their performance on *simple direct* questions versus more complex question types. For instance, stronger models such as GPT-4 and GPT-4o may underperform weaker ones on *simple direct* questions, which primarily assess factual recall. This discrepancy suggests that, for *simple direct* questions, the calibration of LLMs may lead models to exhibit excessive confidence in their internal knowledge, thereby inhibiting their reliance on external documents as the ground truth.

In real-world situations, however, knowledge conflicts typically occur in less established factual domains. As a result, stronger LLMs are expected to demonstrate improved performance in such scenarios due to their ability to better navigate and resolve these conflicts.

Original Paragraph	... The halophiles, which means "salt-loving", live in environments with high levels of salt. They have been identified in the Great Salt Lake in Utah and in the Dead Sea between Israel and Jordan, which have salt concentrations several times that of the oceans. ...
Updated Paragraph	... The halophiles, which means "***pressure***-loving", live in environments with high levels of ***pressure***. They have been identified in the ***Mariana Trench*** in ***western Pacific Ocean***, which have higher ***pressure*** than other environments. ...
Question	In which type of environments do halophiles typically live? A. High Acidity Environments. B. High Salt Environments. C. High Pressure Environments. D. High Sugar Environments.
Original Answer	B. High Salt Environments.
Updated Answer	***C. High Pressure Environments.***

Table 5: An example of hypothetical knowledge update for a question in Biology. The modifications of factual knowledge and contextual information in the paragraph are highlighted in ***red***.

Document Statistics	**Subjects**					**Total**
	Chem.	*Bio.*	*Phys.*	*Geo.*	*Hist.*	
Average Document Length	315.5	588.1	647.3	429.8	277.1	437.3
Num of Documents	291	671	166	471	606	2205

Table 6: Average length (in word counts) and quantities of documents in different subjects.

# Questions (avg. length)	**Question Types**					**Total**
	Simple Direct	*Multi-hop Direct*	*Multi-hop Distant*	*Multi-hop Implicit*	*Distant Implicit*	
Chemistry	73 (12.4)	88 (17.8)	78 (17.9)	40 (17.4)	68 (21.1)	347 (17.3)
Biology	148 (11.1)	170 (16.7)	248 (16.7)	94 (16.7)	213 (20.7)	873 (16.7)
Physics	46 (12.0)	49 (16.7)	49 (15.6)	41 (16.9)	39 (19.9)	224 (16.1)
Geography	141 (12.1)	147 (16.7)	162 (16.8)	96 (17.7)	144 (21.3)	690 (16.9)
History	202 (12.8)	201 (16.1)	171 (17.0)	146 (18.1)	151 (21.6)	871 (16.8)
Total	610 (12.1)	655 (16.7)	708 (16.9)	417 (17.5)	615 (21.0)	3005 (16.8)

Table 7: Average question length (in word counts) and quantities of questions in different subjects and types.

Retrieval Methods	**Subjects**					**Question Types**					**Average**
	chem.	*bio.*	*phys.*	*geo.*	*hist.*	*Sim. Dir.*	*Mul. Dir.*	*Mul. Dis.*	*Mul. Imp.*	*Dis. Imp.*	
BM25	87.90	84.65	74.55	78.26	81.52	79.34	85.04	86.30	75.78	84.23	82.73
Mistral-embed	84.73	77.09	83.48	73.48	80.94	82.13	79.08	79.38	80.81	72.85	78.74
Ada-002	84.44	79.50	82.14	72.75	81.29	82.30	80.76	80.08	80.10	73.01	79.23

Table 8: Retrieval performance (in recall@1 %) of retrieval approaches in different subjects and question types.

Large Language Models	**Subjects**					**Question Types**					**Average**
	chem.	*bio.*	*phys.*	*geo.*	*hist.*	*Sim. Dir.*	*Mul. Dir.*	*Mul. Dis.*	*Mul. Imp.*	*Dis. Imp.*	
Llama3-8b	77.81	75.60	75.45	80.14	78.65	90.33	88.85	89.69	63.55	49.43	77.77
GPT-3.5-turbo	78.93	80.53	82.59	83.04	82.43	92.62	90.53	91.24	71.22	57.72	81.73
GPT-4-turbo	85.30	89.23	86.16	90.00	90.13	95.74	96.18	96.19	81.06	71.71	88.99

Table 9: Performance (in accuracy %) of LLMs in question answering in different subjects and question types.

Model	**Prompting Method**	
	Direct Answer	Locate-and-Answer
Gemini-1.5-flash	75.21	78.54
GPT-3.5-turbo	73.91	81.73
GPT-4o	80.43	87.09

Table 10: Performance (in accuracy %) of LLMs in question answering with different prompting methods.