Retrieval-augmented Generation to Improve Math Question-Answering: Trade-offs Between Groundedness and Human Preference

Zachary Levonian[1] **Chenglu Li**[2] **Wangda Zhu**[3] **Anoushka Gade**[1]
Owen Henkel[4] **Millie-Ellen Postle**[5] **Wanli Xing**[3]
[1]Digital Harbor Foundation [2]University of Utah [3]University of Florida
[4]University of Oxford [5]Rising Academies
zach@digitalharbor.org

Abstract

For middle-school math students, interactive question-answering (QA) with tutors is an effective way to learn. The flexibility and emergent capabilities of generative large language models (LLMs) has led to a surge of interest in automating portions of the tutoring process—including interactive QA to support conceptual discussion of mathematical concepts. However, LLM responses to math questions can be incorrect or mismatched to the educational context—such as being misaligned with a school's curriculum. One potential solution is retrieval-augmented generation (RAG), which involves incorporating a vetted external knowledge source in the LLM prompt to increase response quality. In this paper, we designed prompts that retrieve and use content from a high-quality open-source math textbook to generate responses to real student questions. We evaluate the efficacy of this RAG system for middle-school algebra and geometry QA by administering a multi-condition survey, finding that humans prefer responses generated using RAG, but not when responses are *too* grounded in the textbook content. We argue that while RAG is able to improve response quality, designers of math QA systems must consider trade-offs between generating responses preferred by students and responses closely matched to specific educational resources.

1 Introduction

According to the National Assessment of Educational Progress (NAEP), nearly 40% of high school students lack a basic grasp of mathematical concepts [33], underscoring the need to enhance math education in K-12 environments. One of the most impactful methods to support students' math learning is through math question and answer (QA) sessions tutored by humans. Math QA can be approached with two main focuses: (1) enhancing students' *procedural* fluency with strategies such as step-by-step problem solving for specific math topics and (2) deepening students' *conceptual* understanding through scaffolding such as clarifying math concepts with concrete or worked examples, providing immediate feedback, and connecting math ideas to real-world scenarios [31, 40, 17]. While tutor-led math QA is effective [34], they face challenges such as efficiently allocating tutoring resources, ensuring wide accessibility due to high costs, and scaling up to support a myriad of learners with consistent quality [21, 11].

To address these challenges in math QA, educational researchers have sought AI to build expert systems and intelligent tutoring systems to enhance math learning with procedural practice [39, 4, 2]. However, limited educational research has focused on the potential of AI for improving students' conceptual understanding of math concepts. This study is a preliminary attempt to fill that gap by building the understanding needed to deploy conceptual math QA. We formed a research partnership

NeurIPS'23 Workshop on Generative AI for Education (GAIED).

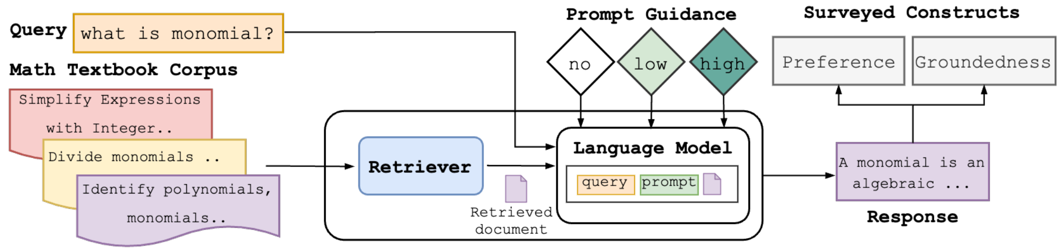


Figure 1: In this paper, we generated responses to math student queries with a retrieval-augmented generation system using one of three prompt guidance conditions. Survey respondents ranked responses by preference and assessed groundedness in the underlying math textbook used as a retrieval corpus.

with the developers of Rori, a WhatsApp-based chatbot math tutor primarily used by low-income middle-school students in Sierra Leone, Liberia, Ghana, and Rwanda.[1]. While Rori uses a chat interface, its pedagogical approach is based on intelligent tutoring systems (ITS) and it adopts a mastery-based learning approach that takes students through procedural lessons based on their abilities. Rori is currently designing for the inclusion of conceptual math QA using LLMs. There have been preliminary efforts to use LLMs in educational settings to scaffold student discussions, to provide feedback [20], to personalize learning experiences through automatic text analysis and generative socio-emotional support [46, 25], and to extend LLMs for many other educational tasks [43].

While the results from these education-related LLM explorations are encouraging, there are ethical considerations when using LLM outputs for math education [20, 35]. A primary concern is *hallucinations*, where LLMs generate answers that sound plausible and coherent but are factually incorrect [12]. Such misleading yet persuasive responses from LLMs could inadvertently instill incorrect conceptual understanding in students. Researchers from the AI community have investigated strategies to mitigate LLM hallucinations (see Ji et al.'s review [19]), with retrieval-augmented generation (RAG) standing out given its effectiveness and flexibility of implementation (e.g., model agnostic) [24, 52]. Conceptually, RAG in an educational context aims to bolster the correctness of LLM-based QA by drawing from external knowledge sources such as syllabi, workbooks, and handouts, such that the LLM's responses are, to various extents, anchored to established learning materials [36]. An interactive student chat backed by RAG offers the promise of both high correctness and faithfulness to materials in a vetted curriculum. Grounding tutoring materials in a student's particular educational context is an important requirement for system adoption [53, 16].

We implemented a RAG system for conceptual math QA (described in sec. 3). To evaluate our RAG system, we started with the problem of designing prompts that produce both the expected tutor-like behavior and responses grounded in the retrieved document. Can we use retrieval-augmented generation and prompt engineering to increase the groundedness of LLM responses? In study 1 (sec. 4), we observe qualitative trade-offs in response quality and the level of guidance provided in the LLM prompt, motivating quantitative study of human preferences. Do humans prefer more grounded responses? In study 2 (sec. 5), we survey preferences for LLM responses at three different levels of prompted guidance, finding that the most-preferred responses strike a balance between no guidance and high guidance. How does retrieval relevance affect response groundedness? In study 3 (sec. 6), we consider the impact of document relevance on observed preferences. Fig. 1 shows an overview of the RAG system and its use for addressing our research questions.

2 Related Work

Intelligent Tutoring Systems (ITSs) are educational technologies designed to provide one-on-one instructional guidance comparable to that of expert human tutors [37]. Structurally, ITSs implement a user interface over a knowledge base with a pedagogical model that determines how the ITS should respond to student inputs [41]. ITSs are traditionally based on iteratively serving procedural lesson content and providing hints in response to student mistakes [48]. ITSs have been shown to be effective

[1]`https://rori.ai`

Table 1: Representative student questions in the 51 Math Nation queries.

Can I get the steps for factoring quadratics	What is the domain and range? How do I find it?
How do I add line segments again??	How do you know if a number is a constant?
what is monomial	How do I multiply fractions???????

as tutors in specific domains such as mathematics and physics [49]. To extend an ITS that currently focuses on procedural fluency with features focused on conceptual understanding [45], we turn to the flexibility and expressive power of LLMs. LLMs have been proposed as useful for supporting a large number of education-related tasks [7, 20].

Despite the potential utility of LLMs for education, there are significant concerns around their correctness and ability to meet students at their appropriate level [20]. LLMs have been used in procedural tutoring and problem-solving systems, with careful prompt engineering used to improve reliability [47]. A more complex approach is using retrieval to augment the LLM prompt in order to improve response quality. For example, the SPOCK system for biology education retrieves relevant textbook snippets when generating hints or providing feedback [44]. Retrieval-augmented generation (RAG) involves retrieving texts from an external corpus relevant to the task and making them available to the LLM [24, 36]. RAG has been used to improve diverse task performance of LLMs [29], either by incorporating retrieved texts via cross-attention [18, 6, 24] or by inserting retrieved documents directly in the prompt [14].[2] We apply RAG in the education domain by using a math textbook as an external corpus and evaluating if RAG leads to responses that are preferred more often by humans and grounded in the textbook content.

3 A retrieval-augmented generation system for math question-answering

To support the development of reliable conceptual question-answering in a math chatbot, we implemented a retrieval-augmented generation system backed by a vetted corpora of math content, e.g. lesson plans, textbooks, and worked examples. RAG cannot provide a benefit during generation if the retrieved documents are not relevant, so we intentionally selected a corpus that will be relevant to many math-related student questions but not to all plausible questions.

OpenStax Prealgebra retrieval corpus We selected a Prealgebra textbook made available by OpenStax [28], segmented by sub-section. The textbook covers whole numbers, functions, and geometry, among other topics.

RAG implementation We adopted a commercially-realistic chatbot context as the underlying LLM, generating all responses with the OpenAI API using model gpt-3.5-turbo-0613 with default temperature settings. We built on our own implementation of RAG [23] that uses a variant of parent retrieval [8]. When a student asks a question, we identify a single relevant section of the textbook using cosine similarity against dense representations of the query and the textbook subsections. We created all representations using OpenAI's text-embedding-ada-002 model [13], an effective dense text embedding model [32]. We released our code and data on GitHub.[3] Additional details in App. A.

4 Study 1: Can we use retrieval-augmented generation and prompt engineering to increase the groundedness of LLM responses?

In using RAG, we hope that system responses will both answer the student's query and reflect the contents of the retrieved document. As the retrieved document cannot be perfectly relevant for all queries, achieving this *groundedness* may require producing inaccurate or otherwise less useful responses. Thus, there is an apparent trade-off between groundedness and the perceived usefulness of the system response. If this trade-off exists, we may want to influence the balance between

[2] A note on terminology: in Lewis et al.'s paper proposing "retrieval-augmented generation", they used the term to refer to an underlying LLM trained or fine-tuned with retrieved documents. The term has come to refer to any combination of LLMs and document retrieval: the method we use in this paper follows the common approach of using in-context learning rather than fine-tuning [22, 27]. A better term for these approaches may be "retrieval-enhanced machine learning" [55], which includes pre-LLM neural models using retrieval e.g. [9].

[3] `https://github.com/DigitalHarborFoundation/rag-for-math-qa`

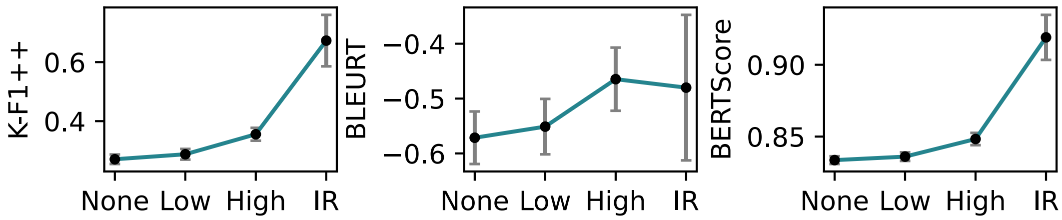


Figure 2: Groundedness for four levels of prompt guidance.

groundedness and usefulness by adjusting the system prompt. This first study tackles a basic question: *can* we influence this balance by engineering the prompt? We now introduce the prompt guidance conditions we used, the queries used for evaluation, and three evaluation metrics.

Guidance conditions Prompt engineering is important for LLM performance [30, 27, 47]. Each guidance condition was selected by iterative, qualitative exploration of prompts given 1-3 sample student questions. While these prompts are unlikely to be "optimal" [51], they produce reasonable outputs. The **No guidance** condition does not use RAG and contains a simple prompt that begins: "You are going to act as a mathematics tutor for a 13 year old student who is in grade 8 or 9 and lives in Ghana. You will be encouraging and factual. Prefer simple, short responses." Other prompts build on this basic instruction set—see App. B. The **Low guidance** prompt adds "Only if it is relevant, examples and language from the section below may be helpful to format your response:" followed by the retrieved document. The **High guidance** prompt instead says "Reference content from this textbook section in your response:". The **Information Retrieval** condition—used only in this first study to demonstrate the shortfalls of automated metrics for conversational responses—says "Repeat the student's question and then repeat in full the most relevant paragraph from my math textbook."

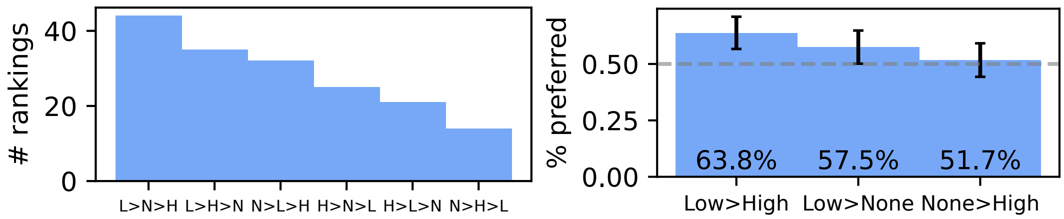
Student queries Math Nation is an online math platform with an interactive discussion board [5]. On this board, students seek help on math-related questions supported by their instructors, paid tutors, and peers. We annotated a random sample of 554 Math Nation posts made by students between October 2013 and October 2021 on boards for Pre-algebra, Algebra 1, and Geometry. We identified 51 factual and conceptual questions that have sufficient context to be answerable; the majority of excluded questions sought procedural help. Representative questions are shown in Table 1.

Evaluation metrics Given the relative novelty of our task, automatically measuring usefulness or correctness is not feasible. However, there is a large body of information retrieval (IR) literature on measuring groundedness of a generated text. We adopt three metrics used in prior work [1, 10, 12, 38]. K-F1++ is a token-level metric that completely ignores semantics, proposed by Chiesurin et al. as more appropriate for conversational QA than Knowledge F1 [10]. BERTScore is a token-level metric that uses RoBERTa-base embeddings to model semantics [56]. BLEURT is a passage-level metric that models semantics using BERT-base fine-tuned on human relevance judgments [42].

Results Fig. 2 shows that metric values on the 51 queries increase across guidance conditions. All confidence intervals are computed at the 95% significance level. These results confirm our basic intuition that groundedness is manipulable with prompt engineering. We do not know if response quality stays the same, increases, or even decreases as groundedness increases, but the results of the IR condition suggest that it *might* decrease: while the token-level metrics indicate that IR is the most grounded condition, its responses include no conversational adaptation to the student's question and so are lower quality in our context. In study 2, we will directly address the questions of response quality and groundedness by surveying humans.

5 Study 2: Do humans prefer more grounded responses?

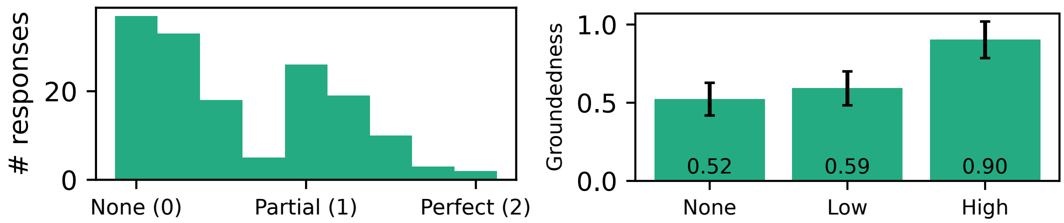
Methods To understand the impact of guidance on human preference for LLM responses, we surveyed 9 educators and designers of education technologies. We selected a comparative (within-subjects) design: with query and response order randomized, respondents ranked from best to worst the responses generated in the None, Low, and High guidance conditions for each query. To determine if the guidance conditions were perceived to be grounded in the retrieved document, we adapted a scale used in prior work as an ordinal None (0), Partial (1), Perfect (2) judgment [1]. Responses were spread across four Qualtrics surveys; all questions received 3-4 responses. The survey is in App. C.



(a) Distribution of all 144 preference rankings.

(b) Pair-wise preferences from all rankings.

Figure 3: Ranked preferences for LLM responses in three guidance conditions: no guidance (N), low guidance (L), and high guidance (H).



(a) Distribution of groundedness across responses in all three guidance conditions, averaged over annotators.

(b) Mean groundedness by guidance condition across all queries and annotators.

Figure 4: Groundedness of the generated responses on an ordinal None (0), Partial (1), Perfect (2) scale.

Results Fig. 3 shows respondent preferences for the three guidance conditions. Responses in the low guidance condition are preferred over responses in the no guidance *and* high guidance conditions. The high and no guidance conditions were statistically indistinguishable. At least two of the guidance conditions significantly differ in groundedness (n=153, one-way ANOVA F(2.0, 99.38)=6.65, p=0.001). We observed substantial inter-rater variation for groundedness ($n = 153$, Krippendorff's α=0.35). Fig. 4 shows that respondents do perceive high guidance responses to be more grounded in the retrieved document than low and no guidance responses. Surprisingly, low guidance responses are not perceived to be significantly more grounded than no guidance responses, suggesting that low guidance responses are preferred for reasons other than their groundedness, a question we will investigate further in study 3.[4]

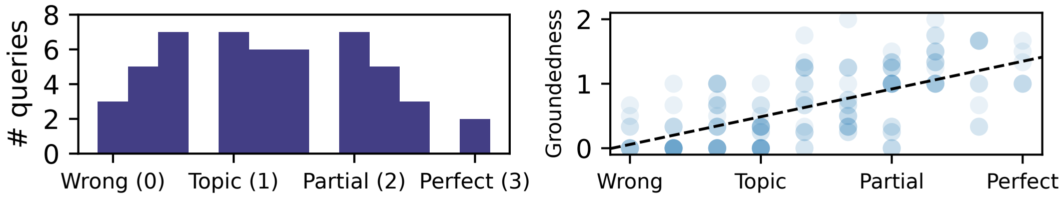
6 Study 3: How does retrieval relevance affect response groundedness?

Methods It may be that responses in the low guidance condition were preferred by survey respondents because the LLM includes content in the retrieved document if it is relevant and omits it if not. To test this hypothesis, three of the authors independently annotated each query and the associated retrieved document for relevance using a four-point ordinal scale used in prior work [15, 3]—see App. D.

Results Inter-rater reliability was generally low ($n = 51$, Fleiss' $\kappa = 0.13$, Krippendorff's $\alpha = 0.40$). For subsequent analysis, we computed the mean relevance of each document across annotators. 70.6% of queries are deemed at least topically relevant, while 33.3% are deemed partially relevant or better; see Fig. 5a for the full distribution. Across all guidance conditions, responses were more likely to be grounded if the retrieved document is relevant (Fig. 5b). However, we observed no significant relationship between relevance and preference (rank). For example, for queries where low guidance responses are preferred over high guidance responses, mean relevance is actually slightly *higher* (diff=0.19, t=-1.45, p=0.15).

Correlation between human annotations and automated metrics Given the results in study 2 suggesting that low guidance responses are not perceived to be more grounded than no guidance responses, we were further interested in possible correlations between perceived groundedness or

[4]Notably, there is no meaningful correlation between the rank of a low guidance response and its perceived faithfulness (Pearson's r=-0.08, p=0.29).



(a) Distribution of retrieved document relevance for all queries, averaged over annotators.

(b) Correlation between perceived groundedness and retrieved document relevance (r=0.56, p<0.001).

Figure 5: Human-annotated relevance of the retrieved document for all 51 queries.

Table 2: Correlation between human annotations and automated groundedness metrics. Pearson's r with p-values Bonferroni-corrected for 12 comparisons. Note: *p<0.05, **p<0.01, ***p<0.001.

Guidance	Faithfulness			Relevance		
	K-F1++	BLEURT	BERTScore	K-F1++	BLEURT	BERTScore
None	0.38	0.33	0.35	0.26	0.34	0.43*
Low	0.47**	0.32	0.61***	0.43*	0.34	0.50**
High	0.50**	0.21	0.39	0.37	0.26	0.50**
Pooled	0.52***	0.33***	0.51***	0.31**	0.30**	0.42***

relevance and the automated groundedness metrics. Table 2 shows modest positive correlations between automated groundedness metrics and human annotations. K-F1++ has the strongest correlation (r=0.52) with groundedness, although the correlation is weaker as guidance decreases.

7 Implications & Future Work

Across three studies, we investigated prompt engineering as a guidance mechanism alongside retrieval-augmented generation to encourage high-quality and grounded responses that are appropriate for students. Our most important finding is that **humans prefer responses to conceptual math questions when retrieval-augmented generation is used, but only if the prompt is not "too guiding".** While RAG is able to improve response quality, we argue that designers of math QA systems should consider trade-offs between generating responses preferred by humans and responses closely matched to specific educational resources. Math QA systems exist within a broader socio-technical educational context; the pedagogically optimal response may not be the one preferred by the student at that time. Chiesurin et al. distinguish between groundedness—when a response is found in the retrieved document—and *faithfulness*—when the response is both grounded and answers the query effectively [10]. Faithfulness is a desirable property for conceptual math QA systems, and we view designing for and evaluating faithfulness as an open problem. Our results show that prompt guidance with RAG is one potential design knob to navigate faithfulness. Future work might improve understanding of faithfulness by building taxonomies based on educational theories of effective tutoring, adapting existing procedural faithfulness metrics (e.g., [1, 12]), and explaining the role of retrieved document relevance (as in our surprising study 3 results finding that relevance was not a meaningful predictor of human preference).

This paper is a preliminary step toward understanding the relationship between groundedness and preference in conceptual math QA systems. Future work must extend beyond single-turn responses to include exploration of follow-up questions [50] and to design for the actual context of use. The most important limitation of this work is that we did not collect preferences directly from middle-school students, although we did use real student questions. Qualitative research of students' preferences should focus not only on correctness but also on factors such as conceptual granularity, curriculuar alignment, and cultural relevance. We were concerned about the ethics of presenting an untested math QA system to students but are now combining insights from these results with the implementation of guard-rails to deploy a safe in-classroom study. Beyond preferences, future math QA systems that use RAG will need to explore the relationship between students' response preferences and actual learning outcomes.

Acknowledgments and Disclosure of Funding

We would like to thank Bill Roberts, Ralph Abboud, and the staff of Rising Academies for their contributions. This work was supported by the Learning Engineering Virtual Institute (LEVI) and funded by the Digital Harbor Foundation.

References

[1] Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2023. Evaluating Correctness and Faithfulness of Instruction-Following Models for Question Answering. http://arxiv.org/abs/2307.16877 arXiv:2307.16877 [cs].

[2] Vincent Aleven, Richard Baraniuk, Emma Brunskill, Scott Crossley, Dora Demszky, Stephen Fancsali, Shivang Gupta, Kenneth Koedinger, Chris Piech, Steve Ritter, Danielle R. Thomas, Simon Woodhead, and Wanli Xing. 2023. Towards the Future of AI-Augmented Human Tutoring in Math Learning. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky (Communications in Computer and Information Science)*, Ning Wang, Genaro Rebolledo-Mendez, Vania Dimitrova, Noboru Matsuda, and Olga C. Santos (Eds.). Springer Nature Switzerland, Cham, 26–31. https://doi.org/10.1007/978-3-031-36336-8_3

[3] Sophia Althammer, Sebastian Hofstätter, Suzan Verberne, and Allan Hanbury. 2022. TripJudge: A Relevance Judgement Test Collection for TripClick Health Retrieval. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. ACM, Atlanta GA USA, 3801–3805. https://doi.org/10.1145/3511808.3557714

[4] Ivon Arroyo, James M Royer, and Beverly P Woolf. 2011. Using an Intelligent Tutor and Math Fluency Training to Improve Math Performance. *International Journal of Artificial Intelligence in Education* 21, 1-2 (2011), 135–152.

[5] Michelle Banawan, Jinnie Shin, Renu Balyan, Walter L. Leite, and Danielle S. McNamara. 2022. Math Discourse Linguistic Components (Cohesive Cues within a Math Discussion Board Discourse). In *Proceedings of the Ninth ACM Conference on Learning @ Scale (L@S '22)*. Association for Computing Machinery, New York, NY, USA, 389–394. https://doi.org/10.1145/3491140.3528320

[6] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022. Improving Language Models by Retrieving from Trillions of Tokens. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR, 2206–2240. https://proceedings.mlr.press/v162/borgeaud22a.html ISSN: 2640-3498.

[7] Andrew Caines, Luca Benedetto, Shiva Taslimipoor, Christopher Davis, Yuan Gao, Oeistein Andersen, Zheng Yuan, Mark Elliott, Russell Moore, Christopher Bryant, Marek Rei, Helen Yannakoudakis, Andrew Mullooly, Diane Nicholls, and Paula Buttery. 2023. On the application of Large Language Models for language teaching and assessment technology. http://arxiv.org/abs/2307.08393 arXiv:2307.08393 [cs].

[8] Harrison Chase. 2023. Parent Document Retriever - LangChain. https://python.langchain.com/docs/modules/data_connection/retrievers/parent_document_retriever

[9] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, Vancouver, Canada, 1870–1879. https://doi.org/10.18653/v1/P17-1171

[10] Sabrina Chiesurin, Dimitris Dimakopoulos, Marco Antonio Sobrevilla Cabezudo, Arash Eshghi, Ioannis Papaioannou, Verena Rieser, and Ioannis Konstas. 2023. The Dangers of trusting Stochastic Parrots: Faithfulness and Trust in Open-domain Conversational Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 947–959. https://doi.org/10.18653/v1/2023.findings-acl.60

[11] Mutlu Cukurova, Madiha Khan-Galaria, Eva Millán, and Rose Luckin. 2022. A Learning Analytics Approach to Monitoring the Quality of Online One-to-One Tutoring. *Journal of Learning Analytics* 9, 2 (2022), 105–120.

[12] Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022. FaithDial: A Faithful Benchmark for Information-Seeking Dialogue. https://doi.org/10.48550/arXiv.2204.10757 arXiv:2204.10757 [cs].

[13] Ryan Greene, Ted Sanders, Lilian Weng, and Arvind Neelakantan. 2022. New and improved embedding model. https://openai.com/blog/new-and-improved-embedding-model

[14] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20, Vol. 119)*. JMLR.org, 3929–3938.

[15] Sebastian Hofstätter, Markus Zlabinger, Mete Sertkan, Michael Schröder, and Allan Hanbury. 2020. Fine-Grained Relevance Annotations for Multi-Task Document Ranking and Question Answering. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 3031–3038. https://doi.org/10.1145/3340531.3412878

[16] Kenneth Holstein, Bruce M. McLaren, and Vincent Aleven. 2017. Intelligent tutors as teachers' aides: exploring teacher needs for real-time analytics in blended classrooms. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference (LAK '17)*. Association for Computing Machinery, New York, NY, USA, 257–266. https://doi.org/10.1145/3027385.3027451

[17] Derek Hurrell. 2021. Conceptual Knowledge or Procedural Knowledge or Conceptual Knowledge and Procedural Knowledge: Why the Conjunction Is Important to Teachers. 46, 2 (2021), 57–71.

[18] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Atlas: Few-shot Learning with Retrieval Augmented Language Models. https://doi.org/10.48550/arXiv.2208.03299 arXiv:2208.03299 [cs].

[19] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. 55, 12 (2023), 1–38.

[20] 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 (April 2023), 102274. https://doi.org/10.1016/j.lindif.2023.102274

[21] Matthew A Kraft and Grace T Falken. 2021. A Blueprint for Scaling Tutoring and Mentoring across Public Schools. 7 (2021), 23328584211042858.

[22] Angeliki Lazaridou, Elena Gribovskaya, Wojciech Jan Stokowiec, and Nikolai Grigorev. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering. (Sept. 2022). https://openreview.net/forum?id=hFCUPkSSRE

[23] Zachary Levonian, Owen Henkel, and Bill Roberts. 2023. llm-math-education: Retrieval augmented generation for middle-school math question answering and hint generation. https://doi.org/10.5281/zenodo.8284412

[24] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, 9459–9474.

[25] Chenglu Li and Wanli Xing. 2021. Natural Language Generation Using Deep Learning to Support MOOC Learners. 31 (2021), 186–214.

[26] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 2356–2362. https://doi.org/10.1145/3404835.3463238

[27] Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2023. Are Emergent Abilities in Large Language Models just In-Context Learning? https://doi.org/10.48550/arXiv.2309.01809 arXiv:2309.01809 [cs].

[28] Lynn Marecek, MaryAnne Anthony-Smith, and Andrea Honeycutt Mathis. 2020. *Prealgebra* (2 ed.). https://openstax.org/details/books/prealgebra-2e

[29] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented Language Models: a Survey. http://arxiv.org/abs/2302.07842 arXiv:2302.07842 [cs].

[30] Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2022. Reframing Instructional Prompts to GPTk's Language. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, Dublin, Ireland, 589–612. https://doi.org/10.18653/v1/2022.findings-acl.50

[31] Judit N Moschkovich. 2015. Scaffolding Student Participation in Mathematical Practices. 47 (2015), 1067–1078.

[32] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. https://doi.org/10.48550/arXiv.2210.07316 arXiv:2210.07316 [cs].

[33] NAEP. 2022. *NAEP Mathematics: National Average Scores*. https://www.nationsreportcard.gov/mathematics/nation/scores/?grade=8

[34] Andre Nickow, Philip Oreopoulos, and Vincent Quan. 2020. The impressive effects of tutoring on prek-12 learning: A systematic review and meta-analysis of the experimental evidence. (2020). Publisher: National Bureau of Economic Research.

[35] 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. Tokyo, Japan. https://ceur-ws.org/Vol-3487/paper4.pdf

[36] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback. http://arxiv.org/abs/2302.12813 arXiv:2302.12813 [cs].

[37] Joseph Psotka, L. Dan Massey, and Sharon A. Mutter (Eds.). 1988. *Intelligent tutoring systems: Lessons learned*. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US. Pages: xxii, 552.

[38] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. `http://arxiv.org/abs/1606.05250` arXiv:1606.05250 [cs].

[39] Steven Ritter, John R Anderson, Kenneth R Koedinger, and Albert Corbett. 2007. Cognitive Tutor: Applied Research in Mathematics Education. 14 (2007), 249–255.

[40] Bethany Rittle-Johnson, Michael Schneider, and Jon R Star. 2015. Not a One-Way Street: Bidirectional Relations between Procedural and Conceptual Knowledge of Mathematics. 27 (2015), 587–597.

[41] P. Sedlmeier. 2001. Intelligent Tutoring Systems. In *International Encyclopedia of the Social & Behavioral Sciences*, Neil J. Smelser and Paul B. Baltes (Eds.). Pergamon, Oxford, 7674–7678. `https://doi.org/10.1016/B0-08-043076-7/01618-1`

[42] Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7881–7892. `https://doi.org/10.18653/v1/2020.acl-main.704`

[43] Jia Tracy Shen, Michiharu Yamashita, Ethan Prihar, Neil Heffernan, Xintao Wu, Ben Graff, and Dongwon Lee. 2021. MathBERT: A Pre-trained Language Model for General NLP Tasks in Mathematics Education.

[44] Shashank Sonkar, Lucy Liu, Debshila Basu Mallick, and Richard G. Baraniuk. 2023. CLASS Meet SPOCK: An Education Tutoring Chatbot based on Learning Science Principles. `http://arxiv.org/abs/2305.13272` arXiv:2305.13272 [cs].

[45] Robert A Sottilare, Arthur Graesser, Xiangen Hu, and Benjamin S Goldberg. 2014. *Design Recommendations for Intelligent Tutoring Systems. Volume 2: Instructional Management*. Technical Report. UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES. 427 pages. `https://apps.dtic.mil/sti/citations/AD1158927` Section: Technical Reports.

[46] Shannon H Sung, Chenglu Li, Guanhua Chen, Xudong Huang, Charles Xie, Joyce Massicotte, and Ji Shen. 2021. How Does Augmented Observation Facilitate Multimodal Representational Thinking? Applying Deep Learning to Decode Complex Student Construct. 30 (2021), 210–226.

[47] Shriyash Upadhyay, Etan Ginsberg, and Chris Callison-Burch. 2023. Improving Mathematics Tutoring With A Code Scratchpad. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Association for Computational Linguistics, Toronto, Canada, 20–28. `https://doi.org/10.18653/v1/2023.bea-1.2`

[48] Kurt VanLehn. 2006. The Behavior of Tutoring Systems. *International Journal of Artificial Intelligence in Education* 16, 3 (Aug. 2006), 227–265.

[49] Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist* 46, 4 (2011), 197–221. `https://doi.org/10.1080/00461520.2011.611369` Place: United Kingdom Publisher: Taylor & Francis.

[50] Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. 2023. MINT: Evaluating LLMs in Multi-turn Interaction with Tools and Language Feedback. `https://doi.org/10.48550/arXiv.2309.10691` arXiv:2309.10691 [cs].

[51] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2023. Large Language Models as Optimizers. `http://arxiv.org/abs/2309.03409` arXiv:2309.03409 [cs].

[52] Kaiyu Yang, Aidan M. Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan Prenger, and Anima Anandkumar. 2023. LeanDojo: Theorem Proving with Retrieval-Augmented Language Models. `http://arxiv.org/abs/2306.15626` arXiv:2306.15626 [cs, stat].

[53] Kexin Bella Yang, Tomohiro Nagashima, Junhui Yao, Joseph Jay Williams, Kenneth Holstein, and Vincent Aleven. 2021. Can Crowds Customize Instructional Materials with Minimal Expert Guidance? Exploring Teacher-guided Crowdsourcing for Improving Hints in an AI-based Tutor. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 119:1–119:24. https://doi.org/10.1145/3449193

[54] Wei Yang, Kuang Lu, Peilin Yang, and Jimmy Lin. 2019. Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 1129–1132. https://doi.org/10.1145/3331184.3331340

[55] Hamed Zamani, Fernando Diaz, Mostafa Dehghani, Donald Metzler, and Michael Bendersky. 2022. Retrieval-Enhanced Machine Learning. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2875–2886. https://doi.org/10.1145/3477495.3531722

[56] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. https://openreview.net/forum?id=SkeHuCVFDr

A Implementation details

We opted to use GPT-3.5 rather than GPT-4 because it reflects a more realistic cost trade-off for the Rori ITS system we are researching. At the time of the study, GPT-3.5 had a context window of 4K tokens; we used up to 3K tokens for document retrieval. The median chapter and sub-section has 5,050 and 185 tokens respectively. We chose dense retrieval both for its popularity in RAG implementations and its dominance on a related retrieval task (not reported here) compared to a strong sparse-retrieval baseline: Pyserini's BM25 implementation [26, 54]).

B Prompts

Prompts used in the various guidance conditions. "{openstax_text}" is replaced with the retrieved text. The None, Low, and High guidance prompts are provided as system prompts, with the student question provided in a separate user prompt. The IR prompt is provided as a user prompt with "{query}" replaced by the student question.

B.1 No guidance (None) prompt

You are going to act as a mathematics tutor for a 13 year old student who is in grade 8 or 9 and lives in Ghana.

You will be encouraging and factual.

Prefer simple, short responses.

If the student says something inappropriate or off topic you will say you can only focus on mathematics and ask them if they have any math-related follow-up questions.

B.2 Low guidance (Low) prompt

You are going to act as a mathematics tutor for a 13 year old student who is in grade 8 or 9 and lives in Ghana.

You will be encouraging and factual.

Only if it is relevant, examples and language from the section below may be helpful to format your response:

===

Table 3: Number of unique queries annotated by each survey respondent.

Annotator	Query Count
A1	30
A2	30
A3	21
A4	21
A5	21
A6	15
A7	15
A8	15
A9	6

{openstax_text}

===

Prefer simple, short responses.

If the student says something inappropriate or off topic you will say you can only focus on mathematics and ask them if they have any math-related follow-up questions.

B.3 High guidance (High) prompt

You are going to act as a mathematics tutor for a 13 year old student who is in grade 8 or 9 and lives in Ghana.

You will be encouraging and factual.

Use examples and language from the section below to format your response:

===

{openstax_text}

===

Prefer simple, short responses.

If the student says something inappropriate or off topic you will say you can only focus on mathematics and ask them if they have any math-related follow-up questions.

B.4 Information Retrieval (IR) prompt

Given a middle-school math student's question, you will identify the most relevant section from a textbook.

Student question: {query}

Repeat the student's question and then repeat in full the most relevant paragraph from my math textbook. If none of them seem relevant, take a deep breath and output the most relevant. Don't say anything else.

Textbook paragraphs:

{openstax_text}

C Ranking & Groundedness Survey

Queries were split into four Qualtrics surveys; three surveys had 15 questions while the fourth had 6 questions. This section gives the exact survey text presented to respondents. 30 queries were annotated three times and the remaining 41 were annotated four times. Table 3 shows per-annotator counts.

C.1 Intro page

This survey will consist of 15 questions. Your progress will save after each question.

Who are you? (Annotator name) _________________

C.2 Query page

(Survey format note: this page is repeated once for each query in the survey.)

C.2.1 Ranking question

Rank these three responses from best to worst response. Consider if the response answers the question and is factually correct.

Student's question:

{query}

	1	2	3
{response1}	◯	◯	◯
{response2}	◯	◯	◯
{response3}	◯	◯	◯

C.2.2 Groundedness question

For each response, does the response or a paraphrase of the response appear anywhere in the following document?

Note: "First response" refers to the first response in the order they appear above, NOT the document you ranked as "1".

The document:

{openstax_text}

None: The response, even paraphrased, does not appear anywhere in the document.

Partial: Part of the response (or a paraphrase of the response) appears in the document.

Perfect: The response (or a paraphrase of the response) appears in the document.

	None	Partial	Perfect
First response	◯	◯	◯
Second response	◯	◯	◯
Third response	◯	◯	◯

C.2.3 Qualitative observation question

Notes/observations, if you want to flag something for later discussion with other annotators or if you spot a survey problem: _________________

D Relevance Survey

Three respondents (A1, A6, and A10) each independently annotated the 51 queries for relevance in separate tabs of a Google Sheet.

D.1 Annotator instructions

Each row contains a middle-school student's question (called the **query**) and an excerpt from a math textbook (called the **document**). Your task is to decide if the document is relevant to the query.

Your options are:

Wrong: The document has nothing to do with the query, and does not help in any way to answer it.

Topic: The document talks about the general area or topic of a query, might provide some background info, but ultimately does not answer it.

Partial: The document contains a partial answer, but you think there should be more to it.

Perfect: The document contains a full answer: easy to understand and it directly answers the question in full.

For readability, I bullet-pointed the paragraphs within each document. It's okay if only one paragraph within the document is relevant: if any paragraph within the document contains a full (or partial) answer, that is sufficient.

Each annotator has their own sheet within this workbook; annotate only within your own sheet, and don't look at others annotations.

D.2 Spreadsheet tab

The annotation sheet had the following columns: query, document, relevance