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

Owen Henkel
University of Oxford
owen.henkel@education.ox.ac.uk

Zach Levonian
Digital Harbor Foundation
zach@levi.digitalharbor.org

Millie -Ellen Postle
Rising Academies
millie.postle@risingacademies.com

Chenglu Li
University of Utah
chenglu.li@utah.edu

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 not match 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 can 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.

Keywords

Question Answering, Retrieval Augment Generation, Math Instruction

1. INTRODUCTION

According to the National Assessment of Educational Progress (NAEP), nearly 40% of high school students lack a basic grasp of mathematical concepts [32], underscoring the need to improve 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 led by human tutors. 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 [17, 30, 40]. While tutor-led math QA is effective [33], it faces challenges such as efficiently allocating tutoring resources, ensuring wide accessibility due to high costs, and scaling up to support a wide range of learners with consistent quality [11, 21].

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 [2, 4, 39] . However, limited educational research has focused on the potential of AI for improving students' conceptual understanding of math concepts. Large Language Models (LLMs) have considerable potential for use within educational environments and provide many potential benefits in the context of mathematical question answering (Q&A). While there has been active research on using LLMs to assist with procedural Q&A, their use in conceptual Q&A is far less explored.

Traditional methods for conceptual Q&A treat it as an information retrieval problem, often employing a text search interface. Consider a spectrum between a standard information retrieval system and a minimally constrained generative text model like an LLM. In the standard information retrieval model, the task is purely to extract information: identify the text excerpt from a selected corpus that is most relevant to the student's question. This method has several advantages, such as the assurance that the extracted information is verified and potentially approved by educators. However, it also presents limitations, such as the inability to tailor responses to a student's ability level or cultural context, and the dependency on the retrieval corpus to contain and retrieve a suitable response to the student's question. This method, while grounded in verified information, may lack relevance in terms of personalization, contextualization, and alignment with the desired curriculum.

In contrast, instruction-fine-tuned LLMs like ChatGPT offer a more adaptable solution. They can generate responses preferred by humans and can adjust their tone and complexity level to match the student's needs. This flexibility makes LLMs appealing for educational applications. However, concerns about inaccuracies and misinterpretations in the generated responses persist. For instance, the information provided might be factually inaccurate whilst appearing authoritative. Even if the information is correct or even "useful", it may be irrelevant to the current classroom lesson.

Accordingly, LLMs offer a high degree of relevance, encompassing personalization, contextualization, and alignment with the preferred curriculum. However, they also pose a significant risk of providing insufficiently grounded information. This issue becomes particularly noticeable in formal learning settings where institutions are held responsible for the accuracy and integrity of the information provided. The optimal solution might be a hybrid model that balances these two needs: grounding the responses in trusted, validated sources relevant to classroom instruction, and relevance by tailoring responses to the specific needs and preferences of the student.

O. Henkel, Z. Levonian, C. Li, and M. Postle. Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 315–320, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12729824

Several strategies could achieve this, with the most promising being retrieval-augmented generation (RAG). RAG combines prompt engineering with a retrieval system. In its basic form, we use a textual prompt to condition the model's generated response. The student's question is incorporated into the prompt until the generated responses meet the desired quality. RAG extends this approach by integrating a retrieval system. The student's query is passed to a retrieval system which identifies relevant texts from a corpus and incorporates these into the prompt before passing it to the large language model.

This study is a preliminary attempt to fill that gap by building the understanding needed to deploy conceptual math QA. 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.

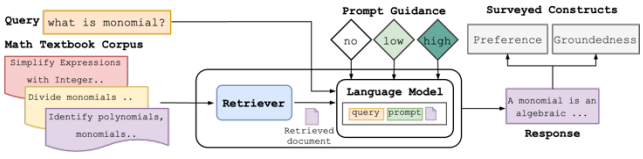


Figure 1: In this paper, we generated responses to math student queries with a retrieval-augmented generation system using one of three prompt guidance conditions. Survey respondents ranked responses by preference and assessed groundedness in the underlying math textbook used as a retrieval corpus.

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 [49]. ITSs have been shown to be effective as tutors in specific domains such as mathematics and physics [50]. 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].

Question Answering - 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 [24, 46], and to extend LLMs for many other educational tasks [43]. One particularly interesting area is Question answering. Question and answers are integral to effective teaching practices [29]. Alongside teacher pedagogical content knowledge, classroom management and climate, the quality of instruction has strong evidence of impact on student outcomes. Much research focuses on teaching students how to ask good questions to improve their higher order thinking and mathematical reasoning, however a wider discussion is to be had as to the types of questions students should ask and the appropriate level of teacher response [10]. Within a typical middle school classroom setting, there are other aggravating factors that negatively impact teacher performance and therefore student learning. These include inadequate time, amount of content and large class sizes that add to teaching loads [15]. Therefore it is likely that teachers do not get to spend the adequate amount of time with each individual student. This is important because we know one-on-one tutoring has the greatest positive impact on student learning [6]. Alongside individual tutoring, in class students prefer seeking help from teachers than their peers [3].

Groundedness vs Relevance in Question Answering - During a mathematics class, the teacher has a working model of all student abilities in that classroom. When a question is asked, that teacher uses not only the materials or curriculum content knowledge that is relevant to that particular lesson, they also draw upon other external forms of knowledge to formulate an intrinsic process of explanation that serves the student who asked the question and the rest of the class [47]. Indeed, if a teacher were to solely repeat verbatim the content from a textbook as an explanation, it is likely that explanation will be ineffective for the student. Mathematics, as a subject, is highly dependent on a strong foundation of prior knowledge and interconnecting schemas [36]. In addition to flexible explanations to student questions based on content knowledge, it is also important that explanations are pitched at an appropriate level. So to answer the question of how important it is to be highly faithful and correct compared to the relevance of the question and the student, in an in-person teaching concept it is far better to explain to students at the level they are at. This is because student questions often reveal knowledge gaps and are an implicit exercise in self-reflection, so QA responses should fill in these missing concepts. For RAG QA, this implies that there will be a preference towards relevance rather than groundedness.

Question Answering with LLMs - LLMs have been used in procedural tutoring and problem-solving systems, with careful prompt engineering used to improve reliability [48]. 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 [23, 35]. RAG has been used to improve diverse task performance of LLMs [27], either by incorporating retrieved texts via cross-attention [7, 18, 23] or by inserting retrieved documents directly in the prompt [14] . 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.

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] . While the results from these education-related LLM explorations are encouraging, there are ethical considerations when using LLM outputs for math education [20, 34]. A primary concern involves hallucinations, instances

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) [23, 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 [16, 53].

3. CURRENT STUDY

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.

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.

Table 1: Representative student questions in the 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???????

OpenStax Prealgebra retrieval corpus - We selected a Prealgebra textbook made available by OpenStax [26], segmented by sub-section. The textbook covers whole numbers, functions, and geometry, among other topics.

RAG implementation - We adopted a 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 an implementation of RAG [22] 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 [31].

3.1 Can retrieval-augmented generation and prompt engineering 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 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 [25, 28, 48]. 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. You will be encouraging and factual. Prefer simple, short responses." Other prompts build on this basic instruction set—see additional details in code/date release. 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, 9, 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 [9]. BERTScore is a token-level metric that uses RoBERTa-base embeddings to model semantics [54]. 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.

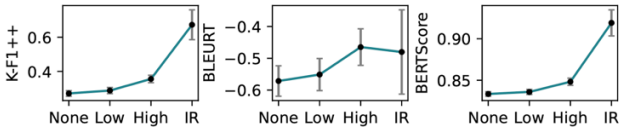
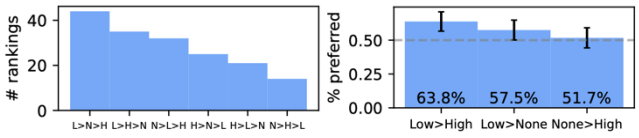


Figure 2: Groundedness for four levels of prompt guidance.

3.2 Do humans prefer more grounded responses?

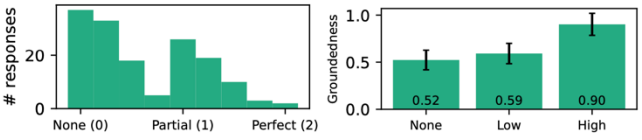
Methods - To understand the impact of guidance on human preference for LLM responses, we surveyed 9 educators and designers of education technologies. We selected a comparative (within-subjects) design: with query and response order randomized, respondents ranked from best to worst the responses generated 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 available in code/date release.

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.



(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).



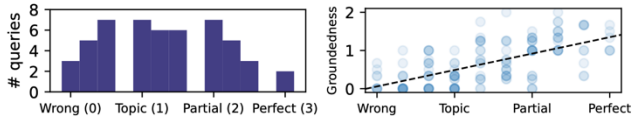
(a) Distribution of groundedness across responses in all three guidance conditions, averaged over annotators. (b) Mean groundedness by guidance condition across all queries and annotators.

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

3.3 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 additional details in code/date release.

Results - Inter-rater reliability was generally low (n = 51, Fleiss' κ = 0.13, Krippendorff's α = 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).



(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.

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 relevance and the automated groundedness metrics. Table 2 shows modest positive correlations be- tween 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.

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***

4. DISCUSSION

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 findings suggest that carefully calibrated prompt guidance within 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. A significant limitation of our study was the absence of direct preference data 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, curricular 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.

5. ACKNOWLEDGMENTS

We would like to thank Ralph Abboud, Nessie Kozhakhmetova, Bill Roberts, Wangda Zhu, Wanli Xing, Anoushka Gade, 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.

All code and data on GitHub at https://github.com/DigitalHarborFoundation/rag-for-math-qa

6. REFERENCES

[1] Adlakha, V., BehnamGhader, P., Lu, X.H., Meade, N. and Reddy, S. 2023. Evaluating Correctness and Faithfulness of Instruction-Following Models for Question Answering. arXiv.

[2] Aleven, V., Baraniuk, R., Brunskill, E., Crossley, S., Demszky, D., Fancsali, S., Gupta, S., Koedinger, K., Piech, C., Ritter, S., Thomas, D.R., Woodhead, S. and Xing, W. 2023. Towards the Future of AI-Augmented Human Tutoring in Math Learning. *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky* (Cham, 2023), 26–31.

[3] Araya, R. and Gormaz, R. 2021. Revealed Preferences of Fourth Graders When Requesting Face-to-Face Help While Doing Math Exercises Online. *Education Sciences*. 11, 8 (Aug. 2021), 429.

[4] Arroyo, I., Royer, J.M. and Woolf, B.P. 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] Banawan, M., Shin, J., Balyan, R., Leite, W.L. and McNamara, D.S. 2022. Math Discourse Linguistic Components (Cohesive Cues within a Math Discussion Board Discourse). *Proceedings of the Ninth ACM Conference on Learning @ Scale* (New York, NY, USA, 2022), 389–394.

[6] Bloom, B. The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring.

[7] Caines, A. et al. 2023. On the application of Large Language Models for language teaching and assessment technology. arXiv.

[8] Chase, H. 2023. Parent Document Retriever - LangChain.

[9] Chiesurin, S., Dimakopoulos, D., Sobrevilla Cabezudo, M.A., Eshghi, A., Papaioannou, I., Rieser, V. and Konstas, I. 2023. The Dangers of trusting Stochastic Parrots: Faithfulness and Trust in Open-domain Conversational Question Answering. *Findings of the Association for Computational Linguistics: ACL 2023* (Toronto, Canada, Jul. 2023), 947–959.

[10] Chin, C. and Osborne, J. 2008. Students' questions: a potential resource for teaching and learning science. *Studies in Science Education*. 44, 1 (Mar. 2008), 1–39.

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

[12] Dziri, N., Kamalloo, E., Milton, S., Zaiane, O., Yu, M., Ponti, E.M. and Reddy, S. 2022. FaithDial: A Faithful Benchmark for Information-Seeking Dialogue. arXiv.

[13] Green, B. and Hu, L. The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning.

[14] Guu, K., Lee, K., Tung, Z., Pasupat, P. and Chang, M.-W. 2020. REALM: retrieval-augmented language model pre-training. *Proceedings of the 37th International Conference on Machine Learning* (Jul. 2020), 3929–3938.

[15] Hayden, T., Stevens and Leko 2018. Teacher Stress: Sources, Effects, and Protective Factors. *Journal of Special Education Leadership*. 31, 2 (2018).

[16] Holstein, K., McLaren, B.M. and Aleven, V. 2017. Intelligent tutors as teachers' aides: exploring teacher needs for real-time analytics in blended classrooms. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (New York, NY, USA, Mar. 2017), 257–266.

[17] Hurrell, D. 2021. Conceptual knowledge or procedural knowledge or conceptual knowledge and procedural knowledge: Why the conjunction is important to teachers. *Australian Journal of Teacher Education (Online)*. 46, 2 (2021), 57–71.

[18] Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S. and Grave, E. 2022. Atlas: Few-shot Learning with Retrieval Augmented Language Models. arXiv.

[19] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A. and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*. 55, 12 (2023), 1–38.

[20] Kasneci, E. et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*. 103, (Apr. 2023), 102274. DOI:https://doi.org/10.1016/j.lindif.2023.102274.

[21] Kraft, M.A. and Falken, G.T. 2021. A blueprint for scaling tutoring and mentoring across public schools. *AERA Open*. 7, (2021), 23328584211042858.

[22] Levonian, Z., Henkel, O. and Roberts, B. 2023. llm-math-education: Retrieval augmented generation for middle-school math question answering and hint generation. Zenodo.

[23] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S. and Kiela, D. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, Dec. 2020), 945.

[24] Li, C. and Xing, W. 2021. Natural Language Generation Using Deep Learning to Support MOOC Learners. *International Journal of Artificial Intelligence in Education*. 31, (2021), 186–214.

[25] Lu, S., Bigoulaeva, I., Sachdeva, R., Madabushi, H.T. and Gurevych, I. 2023. Are Emergent Abilities in Large Language Models just In-Context Learning? arXiv.

[26] Marecek, L., Anthony-Smith, M. and Honeycutt Mathis, A. 2020. *Prealgebra*.

[27] Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., Grave, E., LeCun, Y. and Scialom, T. 2023. Augmented Language Models: a Survey. arXiv.

[28] Mishra, S., Khashabi, D., Baral, C., Choi, Y. and Hajishirzi, H. 2022. Reframing Instructional Prompts to GPTk's Language. *Findings of the Association for Computational Linguistics: ACL 2022* (Dublin, Ireland, May 2022)

[29] Morrison, B. and Evans, S. 2018. University students' conceptions of the good teacher: A Hong Kong perspective. *Journal of Further and Higher Education*. 42, 3 (Apr. 2018), 352–365.

[30] Moschkovich, J.N. 2015. Scaffolding student participation in mathematical practices. *Zdm*. 47, (2015), 1067–1078.

[31] Muennighoff, N., Tazi, N., Magne, L. and Reimers, N. 2023. MTEB: Massive Text Embedding Benchmark. arXiv.

[32] NAEP 2022. NAEP Mathematics: National Average Scores.

[33] Nickow, A., Oreopoulos, P. and Quan, V. 2020. The impressive effects of tutoring on prek-12 learning: A systematic review and meta-analysis of the experimental evidence. (2020).

[34] Nye, B.D. 2015. Intelligent Tutoring Systems by and for the Developing World: A Review of Trends and Approaches for Educational Technology in a Global Context. *International Journal of Artificial Intelligence in Education*. 25, 2 (Jun. 2015), 177–203.

[35] Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu, Z., Chen, W. and Gao, J. 2023. Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback. arXiv.

[36] Poehner, M.E. 2012. The Zone of Proximal Development and the Genesis of Self-Assessment. *The Modern Language Journal*. 96, 4 (Dec. 2012), 610–622.

[37] Psotka, J., Massey, L.D. and Mutter, S.A. eds. 1988. *Intelligent tutoring systems: Lessons learned*. Lawrence Erlbaum Associates, Inc.

[38] Rajpurkar, P., Jia, R. and Liang, P. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Melbourne, Australia, 2018), 784–789.

[39] Ritter, S., Anderson, J.R., Koedinger, K.R. and Corbett, A. 2007. Cognitive Tutor: Applied research in mathematics education. *Psychonomic bulletin & review*. 14, (2007), 249–255.

[40] Rittle-Johnson, B., Schneider, M. and Star, J.R. 2015. Not a one-way street: Bidirectional relations between procedural and conceptual knowledge of mathematics. *Educational Psychology Review*. 27, (2015), 587–597.

[41] Sedlmeier, P. 2001. Intelligent Tutoring Systems. *International Encyclopedia of the Social & Behavioral Sciences*. N.J. Smelser and P.B. Baltes, eds. Pergamon. 7674–7678.

[42] Sellam, T., Das, D. and Parikh, A. 2020. BLEURT: Learning Robust Metrics for Text Generation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online, Jul. 2020), 7881–7892.

[43] Shen, J.T., Yamashita, M., Prihar, E., Heffernan, N., Wu, X., Graff, B. and Lee, D. 2021. MathBERT: A Pre-trained Language Model for General NLP Tasks in Mathematics Education. (2021).

[44] Sonkar, S., Liu, L., Mallick, D.B. and Baraniuk, R.G. 2023. CLASS Meet SPOCK: An Education Tutoring Chatbot based on Learning Science Principles. arXiv.

[45] Sottilare, R.A., Graesser, A., Hu, X. and Goldberg, B.S. 2014. *Design Recommendations for Intelligent Tutoring Systems. Volume 2: Instructional Management*. UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES.

[46] Sung, S.H., Li, C., Chen, G., Huang, X., Xie, C., Massicotte, J. and Shen, J. 2021. How does augmented observation facilitate multimodal representational thinking? Applying deep learning to decode complex student construct. *Journal of Science Education and Technology*. 30, (2021), 210–226.

[47] Tofade, T., Elsner, J. and Haines, S.T. 2013. Best Practice Strategies for Effective Use of Questions as a Teaching Tool. *American Journal of Pharmaceutical Education*. 77, 7 (Sep. 2013), 155.

[48] Upadhyay, S., Ginsberg, E. and Callison-Burch, C. 2023. Improving Mathematics Tutoring With A Code Scratchpad. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (Toronto, Canada, Jul. 2023), 20–28.

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

[50] Vanlehn, K. 2011. The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist*. 46, 4 (Oct. 2011), 197–221. DOI:https://doi.org/10.1080/00461520.2011.611369.

[51] Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q.V., Zhou, D. and Chen, X. 2023. Large Language Models as Optimizers. arXiv.

[52] Yang, K., Swope, A.M., Gu, A., Chalamala, R., Song, P., Yu, S., Godil, S., Prenger, R. and Anandkumar, A. 2023. LeanDojo: Theorem Proving with Retrieval-Augmented Language Models. arXiv.

[53] Yang, K.B., Nagashima, T., Yao, J., Williams, J.J., Holstein, K. and Aleven, V. 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 (Apr. 2021), 119:1-119:24.

[54] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q. and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. (2020).