Designing Safe and Relevant Generative Chats for Math Learning in Intelligent Tutoring Systems

Zachary Levonian
Digital Harbor Foundation
Baltimore, United States
zach@levi.digitalharbor.org

Owen Henkel
University of Oxford
Oxford, United Kingdom
chenglu.li@utah.edu

Chenglu Li
University of Utah
Salt Lake City, United States
chenglu.li@utah.edu

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

Large language models (LLMs) are flexible, personalizable, and available, which makes their use within Intelligent Tutoring Systems (ITSs) appealing. However, their flexibility creates risks: inaccuracies, harmful content, and non-curricular material. Ethically deploying LLM-backed ITSs requires designing safeguards that ensure positive experiences for students. We describe the design of a conversational system integrated into an ITS that uses safety guardrails and retrieval-augmented generation to support middle-grade math learning. We evaluated this system using red-teaming, offline analyses, an in-classroom usability test, and a field deployment. We present empirical data from more than 8,000 student conversations designed to encourage a growth mindset, finding that the GPT-3.5 LLM rarely generates inappropriate messages and that retrieval-augmented generation improves response quality. The student interaction behaviors we observe provide implications for designers—to focus on student inputs as a content moderation problem—and implications for researchers—to focus on subtle forms of bad content and creating metrics and evaluation processes.[1]

Keywords: large language models, intelligent tutoring systems, safety, system design

[1]Code and data are available at https://www.github.com/DigitalHarborFoundation/chatbot-safety and https://www.github.com/DigitalHarborFoundation/rag-for-math-qa

1. INTRODUCTION

According to the National Assessment of Educational Progress (NAEP), nearly 40% of high school students lack a basic grasp of mathematical concepts (NAEP, 2022), underscoring the need to improve math education in K-12 environments. One of the most impactful methods to support students' math learning is through conversation with teacher or tutors. These conversations can help enhance students' procedural fluency with strategies such as step-by-step problem solving for specific math topics or by 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 (Hurrell, 2021; Moschkovich, 2015; Rittle-Johnson et al., 2015). Conversations can also be effective at providing meta-cognitive support to students independent of specific math concepts, such as by encouraging students to adopt a growth mindset (Karumbaiah et al., 2017). While teacher-led conversation is effective (Nickow et al., 2020), 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 (Kraft and Falken, 2021; Cukurova et al., 2022).

The capabilities of large language models (LLMs) have led to a surge of interest in addressing these scaling challenges by applying LLMs for automated tutoring, personalized learning, and adaptive assessment (Kasneci et al., 2023; Caines et al., 2023). A particularly promising application of LLMs is integration with Intelligent Tutoring Systems (ITSs), as they can combine the structured pedagogical processes and vetted curricula of ITSs and the flexibility and personalization enabled by conversational interfaces (Upadhyay et al., 2023; Hobert and Wolff, 2019; Khosrawi-Rad et al., 2022).

Integrating LLMs into ITSs enables answering student questions, summarizing concepts, creating customized hints, and recontextualizing learning materials (Pardos and Bhandari, 2023; Sonkar et al., 2023). While there has been active research on using LLMs to assist with procedural practice, their use in conceptual and socio-emotional questions is far less explored. Traditional methods for question answering 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 some 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, the use of LLMs in educational applications also raises concerns regarding potential risks, including the generation of toxic language, implicit biases, and inaccurate information, as well as inappropriate use by students (Liang et al., 2021; Navigli et al., 2023; Baker and Hawn, 2022; Yan et al., 2024). These risks become particularly important when designing educational applications that directly interact with students, e.g., via a chat interface, necessitating a focus

on the safety and accuracy of model-generated responses to students.

Recent advancements in LLMs have led to improvements in mitigating some of the most distressing behaviors of early generations, such as toxicity, wildly inaccurate information, and discussions of illegal or taboo topics (OpenAI, 2023; Tao et al., 2024). While this progress is welcome, it has also revealed a range of more subtle potential problems. For instance, small hallucinations (e.g., confusing "2πr" with "πr^2") may lead to persistent misconceptions (Murphy and Alexander, 2013). Additionally, younger students might be more likely to anthropomorphize models and develop emotionally charged relationships with them (Girard and Johnson, 2010; Goldman et al., 2023), and models tend to present an "average" view of the Anglophone internet which might not be appropriate in certain cultural contexts (Xu et al., 2024; Agiza et al., 2024).

Less discussed is how models should handle inappropriate or potentially offensive student inputs, as well as honest questions on politically or culturally sensitive topics. For example, if a student addresses an LLM application using profane language, should the model ignore the profanity and proceed, ask the student to stop using such language, or request that the student rephrase the question? Similarly, if a student asks an honest question about a potentially charged political topic (e.g., "Is it okay to get pregnant before you are married?"), should the model provide a standard "it depends" answer, ignore the question, or inform the student that they cannot discuss the topic? Perhaps most seriously, if a student discloses some sort of trauma or abuse they have suffered, how should the model respond?

While these are complex questions, they are also ones that teachers and tutors deal with regularly (Falkiner et al., 2017; Beck et al., 1994; Haney, 2004; Berger et al., 2022). Deciding how to respond to inappropriate or provocative student questions is a classic challenge of classroom management (Sabornie and Espelage, 2022; Marzano, 2005), carefully choosing how to address and explain sensitive topics is a fraught area for nearly all teachers (Levin and Nolan, 2002; Falkiner et al., 2020), and handling sensitive student disclosures is such an important question that most school systems have codified mandatory reporting rules for teachers that specify which types of student disclosures must be reported to school leadership, mental health professionals, or law enforcement (Goldman, 2007).

While LLMs offer a high degree of relevance—encompassing personalization, contextualization, and alignment with the preferred curriculum—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 while still tailoring the responses to the specific needs and preferences of the student.

In this paper, we describe a system we designed for safeguarding student conversations within an ITS and empirical data from a field deployment of that system with usage from more than 8,000 students. We formed a research collaboration with the developers of Rori, a WhatsApp-based chatbot math tutor. Rori is used primarily by low-income middle-school students in Sierra Leone, Liberia, Ghana, and Rwanda both in classroom settings and at home for math skills practice (Henkel et al., 2024). As do many math-focused ITSs, Rori primarily provides structured, procedural math practice. We incorporated LLMs in Rori in two ways: (1) a *structured conversation* for new users that teaches students about growth mindset before math skill practice begins, and (2) an open-ended *Q&A* for answering student's conceptual math questions as a tutor might.

We provide evidence for the utility and practicality of integrating LLMs in ITSs. In Study 1 (sec. 4), we assess the usability and ethical acceptability of structured conversations in a classroom usability test and in a field deployment. In Study 2 (sec. 5), we assess the usefulness of retrieval-augmented generation for answering student's conceptual questions. The empirical evidence from these two studies provides implications for designers—to focus on student inputs as a content moderation problem—and implications for researchers—to focus on subtle forms of bad content and creating metrics and evaluation processes appropriate for the ITS context. This journal article is a direct extension of (Henkel et al., 2024) and (Levonian and Henkel, 2024). Ethically incorporating LLMs into education systems is a research problem that intersects AI ethics, socio-emotional learning, and equitable education. This study contributes to this research problem by incorporating LLMs into ITSs with safeguards to promote safe interactions and socio-emotional learning for diverse student populations.

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 (Psotka et al., 1988). Structurally, ITSs implement a user interface over a knowledge base with a pedagogical model that determines how the ITS should respond to student inputs (Sedlmeier, 2001). ITSs are traditionally based on iteratively serving procedural lesson content and providing hints in response to student mistakes (VanLehn, 2006). ITSs have been shown to be effective as human tutors in specific domains such as mathematics and physics (VanLehn, 2011). To extend an ITS that currently focuses on procedural fluency with features focused on conceptual understanding and meta-cognitive skills (Sottilare et al., 2014), we turn to the flexibility and expressive power of Large Language Models (LLMs). LLMs have been proposed as useful for supporting a large number of education-related tasks (e.g., Caines et al. 2023; Kasneci et al. 2023). There have been preliminary efforts to use LLMs in educational settings to scaffold student discussions, such as by providing feedback (Kasneci et al., 2023; Henkel et al., 2024), personalizing learning experiences through automatic text analysis and generative socio-emotional support (Sung et al., 2021; Li and Xing, 2021), and extending LLMs for other math education applications (Shen et al., 2021).

Despite the potential utility of LLMs for education, there are significant concerns around their correctness and ability to meet students at their appropriate level (Kasneci et al., 2023). LLMs have been used in procedural tutoring and problem-solving systems, with careful prompt engineering used to improve reliability (Upadhyay et al., 2023). 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 (Sonkar et al., 2023). Retrieval-augmented generation (RAG) involves retrieving texts from an external corpus relevant to the task and making them available to the LLM (Lewis et al., 2020; Peng et al., 2023). RAG has been used to improve diverse task performance of LLMs (Mialon et al., 2023), either by incorporating retrieved texts via cross-attention (Izacard et al., 2024; Borgeaud et al., 2022; Lewis et al., 2020) or by inserting retrieved documents directly in the prompt (Guu et al., 2020).[2] In Study 2 (sec. 5), we apply

[2]A note on terminology: (Lewis et al., 2020) proposed "retrieval-augmented generation" to refer to an underlying LLM trained or fine-tuned with retrieved documents. The term has come to refer to any combination of

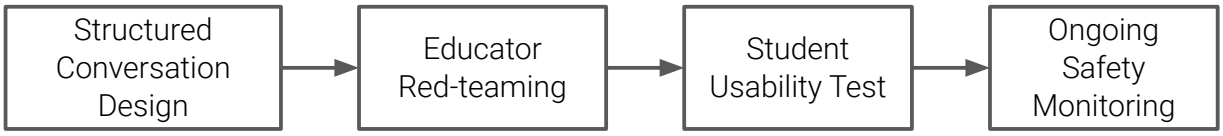


Figure 1: Designing for safety: our process.

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.

Even when LLMs produce "correct" outputs, use of LLMs in ITSs presents risks. Initial evidence suggests that LLMs might harm student learning or engagement (Bastani et al., 2024; Lehmann et al., 2024; Nie et al., 2024). Students frequently use systems in unintended ways (Baker, 2007), and unfortunately LLMs produce unexpected responses in response to unusual inputs (Hendrycks et al., 2022; Gabriel et al., 2024). So LLMs might generate harmful or inappropriate messages in response to student use (Dinan et al., 2021), compromising student safety. One approach to decreasing these safety risks is not presenting LLM outputs directly to students and instead presenting them to an educator who can vet the generated contents, presenting them to students as appropriate (Wang et al., 2024; Demszky et al., 2024). In this paper, we focus on presenting LLM outputs directly to student users of ITSs. The direct exposure risks to humans from interacting with LLMs—especially in an iterative, conversational mode—are generally underexplored (Fischer, 2023). Our understanding of how students use and benefit from LLMs is still preliminary, and many students report unhelpful interactions with LLMs (Maiti and Goel, 2024; Lieb and Goel, 2024). This lack of knowledge combined with the potential for LLMs to generate harmful responses makes deploying LLMs for use by students risky. We reduce that risk by exploring students' iterative interaction with an LLM during a structured conversation (Study 1, sec. 4) and during open-ended Q&A (Study 2, sec. 5).

3. System Design

To design a safe generative chat experience, we implemented a system on the basis of educator feedback and through multiple phases of evaluation as shown in Figure 1. We release open-source implementations of the moderation[3] and retrieval-augmented generation[4] components on GitHub. We deployed this system inside Rori, an ITS designed around "micro-lessons" that explain a math concept alongside multiple-choice or free response practice questions. The system extension we designed for these two studies consists of two modules. The first module (sec. 3.1) is a structured conversation teaching a student about the concept of a growth mindset, presented to students before they begin math practice. The second module (sec. 3.2) supports open-ended Q&A, presented to students when they ask questions during math practice.

LLMs and document retrieval: the method we use in this paper follows the common approach of using in-context learning rather than fine-tuning (Lazaridou et al., 2022; Lu et al., 2023). A better term for these approaches may be "retrieval-enhanced machine learning" (Zamani et al., 2022), which includes pre-LLM neural models using retrieval (e.g., Chen et al. 2017).

[3] https://github.com/DigitalHarborFoundation/chatbot-safety

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

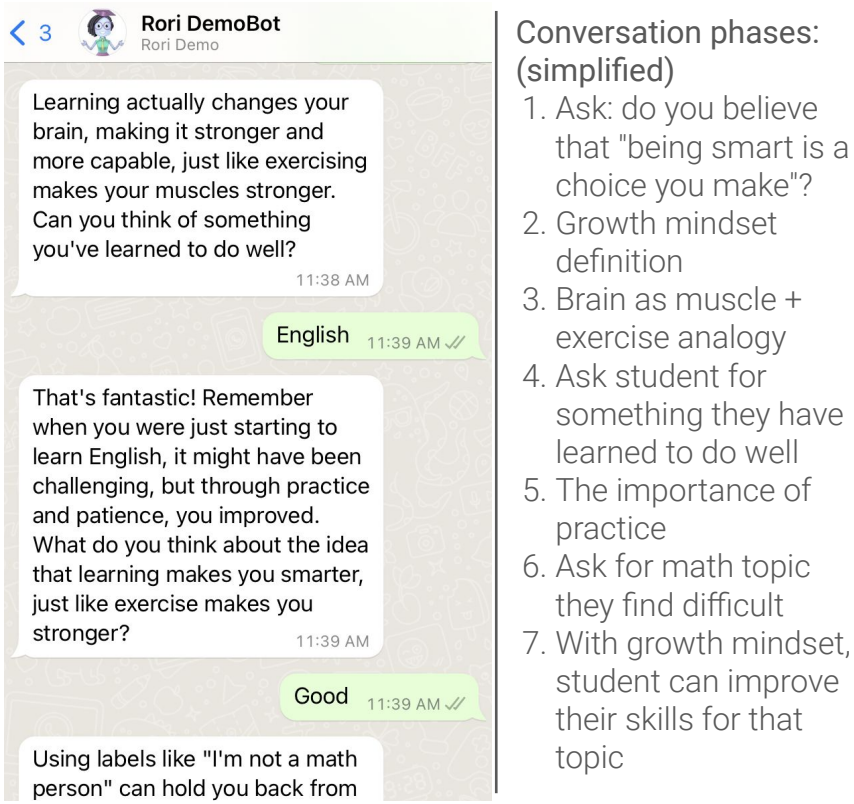


Figure 2: A chat excerpt from the Rori WhatsApp interface and a simplified view of the conversation phases.

3.1. Designing a Structured Conversation

We chose to implement a generative chat for encouraging a growth mindset, an approach linked to positive educational outcomes, including in mobile learning contexts (Karumbaiah et al., 2017; Yeager et al., 2016; Kizilcec and Goldfarb, 2019). We used a prompting approach that moves the conversation through multiple phases: introducing the concept of a growth mindset, asking the student to reflect on a time that practice has helped them, and identifying a specific math skill that they want to practice. The system initiates the conversation with the message "Do you agree with the statement 'Being smart is a choice you make, not the way you are'?" and moves the student through various conversational phases, as shown in Figure 2. During the conversation, we detect standard navigation keywords (e.g. "menu") to navigate away from the conversation and on to math skills practice. We limited the total conversation length—a max of 8 turns during the usability test and 10 during the field deployment—to decrease the chance of major digressions and to reduce any student frustration. By designing the conversation as system-initiated rather than student-initiated and ending each system message with a question, we provide structure that keeps the conversation flowing and focused on growth mindset.

3.1.1. Safety Guardrails

To ensure students have a safe experience during the conversation, we implemented guardrails that would redirect or end the conversation. Given the structured nature of the conversation, the baseline risk of an unsafe experience is relatively low. However, learning systems have pedagog-

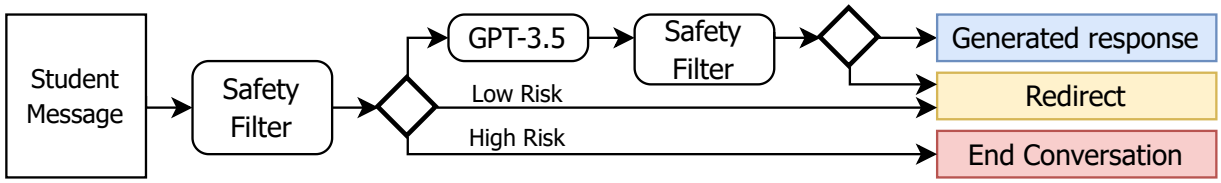


Figure 3: The generative chat moderation system. A Safety Filter—implemented using the OpenAI Moderation API—is applied to both student messages and GPT-3.5-generated responses. Messages that trigger the safety filter produce one of two moderation actions: an attempted redirect, where the student is prompted to try again (for low risk messages), or ending the conversation immediately (for high risk messages).

ical goals that require ITS design to account for student misuse (Aleven and Koedinger, 2001; Baker et al., 2004). Students can be disruptive (Good and Brophy, 2008; Emmer and Stough, 2001), including in online learning environments (Rodrigo et al., 2013). We have previously noticed that some students anthropomorphize Rori or send messages that align with politeness standards around interaction with educators (Graesser et al., 1995). In this context, Rori has an implicit socialization role, similar to a human tutor (Good and Brophy, 2008). Thus, the system must identify student and system-generated messages that compromise the system's pedagogical goals.

Each student and system message is passed through a safety filter that determines how the system will respond to the student. Figure 3 demonstrates the final design. The safety filter consists of (1) a word list and (2) a statistical moderation model. The word list—consisting only of unambiguous curse words—is applied first. While a word list is rigid and inflexible, we chose to include it because it is easier for educators and parents to reason about than a statistical model (Jhaver et al., 2023). The statistical model we used was the OpenAI (2024) moderation API, which at the time of the study predicted the presence of five high-level content categories and six sub-categories (these categories are listed in Table 4). Each message is given a score between 0 and 1 reflecting how likely that message is to contain content in that category. We set the per-category thresholds for which we would take system action based on the red-teaming exercise.

3.1.2. System Moderation Actions

Based on the assessed risk of the message, we took one of two moderation actions in response to student messages. We classified self-harm, sexual/minors, and the two /threatening subcategories as high risk messages and the rest as low risk. In response to low risk messages, we drop the student's most recent message from the prompted context and ask them to continue the conversation with a more appropriate message. In response to high risk messages, we end the conversation immediately with the message: "That sounds like a serious topic, and a real person needs to look at this. They might try to contact you to check on you. Until someone has reviewed this, Rori will not reply." We discuss the need for more sophisticated moderation responses in the Discussion.

3.1.3. Educator Red-teaming

To evaluate the acceptability of the conversation design and the safety guardrails, we conducted an asynchronous red-teaming exercise. There is considerable variation in red-teaming exercises (Feffer et al., 2024); the purpose of our exercise was to qualitatively assess the effectiveness of the safety guardrails and to quantitatively set initial per-category moderation thresholds. We recruited 17 Rising Academies educators and system designers to adversarially probe the conversation design. Adversarial inputs varied widely, but were generally off-topic, disruptive, abusive, or graphic. Across 57 conversations, we received negative feedback on 39 messages that should have been flagged, setting the thresholds appropriately. After small tweaks to the prompts, we observed no obviously negative conversational experiences. We return to the topic of subtly negative experiences in the Discussion, but we determined there to be minimal risk in proceeding with a full usability assessment with students.

3.1.4. Monitoring

To ensure the safety of Rori student users, we designed a continual monitoring procedure. We implemented data dashboards to review the most recent and the riskiest conversations. Messages flagged as high risk generate an email alert to an internal team. We designed a basic reporting protocol for use with student users in the event of particular sensitive disclosures, e.g., sexual abuse or suicidal thoughts. We have not yet had cause to use this reporting protocol.

While the structured conversation provides a controlled environment for introducing important concepts like growth mindset, we recognized the need for more flexible interaction to address students' specific math questions. So, we designed an open-ended Q&A system to provide grounded, relevant responses to student queries.

3.2. Designing for Conceptual Q&A

Educational researchers have used machine learning to build expert systems and intelligent tutoring systems to enhance math learning with procedural practice (Ritter et al., 2007; Arroyo et al., 2011; Aleven et al., 2023). Using LLM outputs to focus on *conceptual* Q&A is appealing, although there are significant ethical considerations (Kasneci et al., 2023; Nye et al., 2023). A primary concern for math Q&A is hallucination, where an LLM generates an answer that sounds plausible and coherent but is factually incorrect (Dziri et al., 2022). Such misleading yet persuasive responses from LLMs could inadvertently instill incorrect conceptual understanding in students. These concerns are particularly relevant in open-domain, unstructured Q&A. Researchers from the machine learning community have investigated strategies to mitigate LLM hallucinations (see review by Ji et al. 2023), with retrieval-augmented generation (RAG) standing out given its effectiveness and flexibility of implementation (e.g., it can be used with any LLM) (Lewis et al., 2020; Yang et al., 2023). Conceptually, RAG in an educational context aims to bolster the correctness of LLM-based Q&A 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 (Peng et al., 2023). 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 (Yang et al., 2021; Holstein et al., 2017).

To support the development of reliable conceptual Q&A in a math chatbot, we implemented a retrieval-augmented generation system backed by a vetted corpora of math content, including

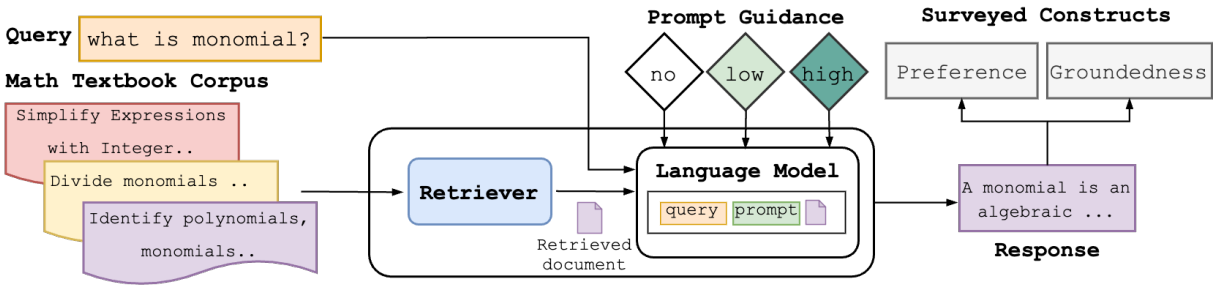


Figure 4: To support conceptual Q&A, we generated responses to math student queries with a retrieval-augmented generation system using one of three prompt guidance conditions. Survey respondents ranked responses by preference and assessed groundedness in the underlying math textbook used as a retrieval corpus. We evaluate this system in Study 2 (sec. 5).

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.

3.2.1. Retrieval Corpus: OpenStax Prealgebra textbook

We selected a prealgebra textbook made available by OpenStax (Marecek et al., 2020), segmented by sub-section. The textbook covers whole numbers, functions, and geometry, among other topics.

3.2.2. Retrieval-augmented Generation 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 (Levonian et al., 2023) that uses a variant of parent retrieval (Chase, 2023). 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 (Greene et al., 2022), an effective dense text embedding model (Muennighoff et al., 2023). Fig. 4 shows an overview of the RAG system. Additional details in App. 8.1. We will evaluate our RAG system in Study 2 (sec. 5).

3.3. Evaluation

To evaluate the effectiveness and safety of our system design, we conducted two interconnected studies. In Study 1 (sec. 4), we evaluate the structured conversation design with a student usability test and a field deployment. In Study 2 (sec. 5), we evaluate open-ended conceptual Q&A capabilities on a dataset of student questions.

4. Study 1: Structured Conversation

Study 1 evaluated the safety of the structured conversation about growth mindset in two phases. The first phase (sec. 4.1) was an in-classroom usability test with 109 students. The second phase

(sec. 4.2) was a field deployment with more than 8,000 students. Summary counts are shown in Table 1.

Table 1: Counts of students, conversations, and messages in Study 1.

	Students	Conversations	Messages
Usability Test	109	252	3,722
Field Deployment	8,168	8,755	126,278

4.1. Phase 1: Student Usability Test

Table 2: Student conversation ratings during Study 1.

Rating	none	☆☆☆☆☆	☆☆☆☆	☆☆☆	☆☆	☆
# conversations	125	126	4	5	2	5

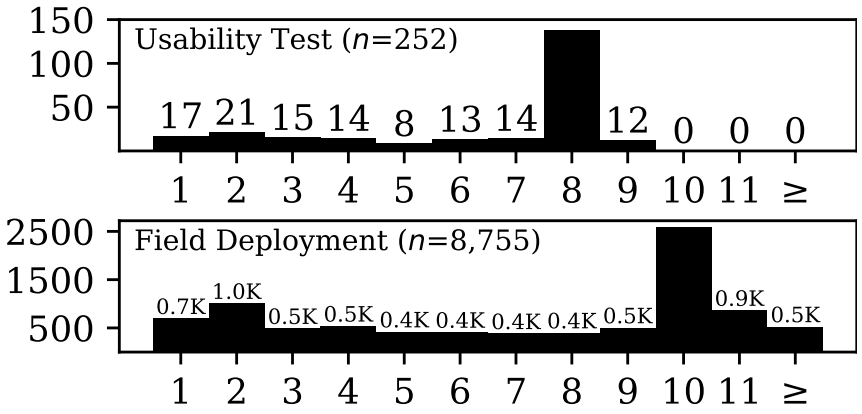


Figure 5: Conversation length (as number of student messages) for all conversations. Completion rate was higher during the usability test (59.5%) than the field deployment (38.9%).

In December 2023, 109 in-school students across six total classrooms were instructed to use the growth mindset generative chat during a regularly-scheduled study hall using Rori for math skills practice (Henkel et al., 2024). 252 conversations occurred between December 13th and 15th. 60% of the conversations were completed; the distribution of conversation lengths is shown in Figure 5.[5]

At the end of the conversation, we asked students to rate the conversation from one to five stars.[6] The distribution of ratings is shown in Table 2. Of the rated conversations, 16 conversations (6.3%) were rated less than five stars. Qualitative investigation of those 16 low-rated

[5]Due to a bug that under-counted student messages, some conversations continued an extra turn, for a total of 9.

[6]Feedback request message: "Thank you for your time! How much did you like the conversation?" A response model labeled "Give us some ☆s!" has quick-reply buttons.

conversations reveals no clear difference between those and 5-star conversation; student messages in low-rated conversations were non-significantly more likely to be single-word responses (75.4% in low-rated conversations vs 65.4% in five-star ones, χ^2=0.82, d.f.=1, p=0.36).

No student or GPT-3.5 student messages were flagged by the safety filter. In fact, most GPT-3.5 and student messages received low moderation scores across all categories. Table 3 shows summary statistics for the highest score received across all categories: the highest-scoring GPT-3.5 message received a score of 0.010 ("Oh, it seems like you might not understand the question. Let me rephrase it. Do you think that being smart is something that you can choose to be, rather than something that you are born with?"), while the highest-scoring student message received a score of 0.045 (a typo).

Table 3: Highest and 99th percentile of the OpenAI moderation scores observed during the two studies. The highest possible value is 1.

	Usability Test		Field Deployment	
Source	Q99	Max	Q99	Max
GPT-3.5	0.000	0.010	0.003	0.044
Student	0.002	0.045	0.030	0.989

It may be that the moderation API's implicit values diverge from our own, such that false negatives occur and harmful student messages are not flagged. To check, we randomly sampled 100 student conversations, finding no false negatives. Qualitatively, while some student messages were playful or inappropriate in ways that would likely trigger a response from a human tutor, we found our prompt for GPT-3.5 effective at producing appropriate redirections back to the current topic.

Taken together, these results suggest that the semi-structured growth mindset conversation is acceptable for broader use. Critically, the conversation design was effective at preventing messages that would trigger the safety filter: we identified no obviously unacceptable student messages. The promising results from our initial usability test encouraged us to expand our investigation to a larger, more diverse user base. In phase 2, we deploy the growth mindset conversation in a real-world setting to assess its performance and safety at scale.

4.2. Phase 2: Field Deployment

The growth mindset conversation was deployed publicly on Feburary 13, 2024 for non-school users of Rori and incorporated as a component of the on-boarding process before math skills practice begins. We analyzed the 126,278 messages between the feature launch and May 1, 2024.

4.2.1. Did GPT-3.5 Generate Objectionable Outputs?

No. Quantitatively, the highest-scoring system message produced received a score of 0.044. During continual monitoring, the researchers annotated GPT-3.5 messages and determined none of them to be objectionable. The most controversial messages were those generated in response to student's objectionable messages, which we discuss in the next sections.

Table 4: OpenAI moderation scores by category for the 54,384 student messages sent during the field deployment. In addition to the 99th percentile and maximum observed score over all student messages, we show the number of messages with a score greater than 0.1 and greater than 0.5.

Category	Q99	Max	$n \geq 0.1$	$n \geq 0.5$
Harassment	0.011	0.989	141	36
Sexual	0.012	0.914	28	5
Hate	0.002	0.524	3	1
Violence	0.001	0.959	2	1
Self-harm/intent	0.001	0.743	1	1
Self-harm	0.001	0.531	1	1
Harassment/threatening	0.000	0.451	1	0
Hate/threatening	0.000	0.087	0	0
Violence/graphic	0.000	0.081	0	0
Self-harm/instructions	0.000	0.072	0	0
Sexual/minors	0.007	0.024	0	0

4.2.2. Did Students Write Objectionable Messages?

Yes, but not very much. 0.31% of student messages received a score in any moderation category of at least 0.1. Fewer than 8 in 10000 messages were flagged. Table 4 summarizes the moderation scores per-category. The most common negative messages were harassing or sexual. Only one message was flagged as high risk. After investigation by the team, it was determined to be a false positive by the OpenAI moderation model—the message should have been classified as low risk, as it contained violent language that merited corrective action but did not evidence self-harm. From an investigation of the 27 conversations with flagged messages, all flagged messages were determined to merit corrective action.

4.2.3. Did GPT-3.5 Respond Appropriately?

We investigated the messages generated in response to student messages that were near the safety filter thresholds but remained unflagged. 48 unflagged conversations contained a message with a moderation score of 0.1 or higher. 40 of these conversations included at least one student message that warranted caution or a corrective statement from the system response, and we deemed the GPT-3.5-generated responses to be appropriately corrective in 37 of those cases. In 3 cases, the generated response ignored or equivocated when a corrective message would have been warranted. This is a subtler form of bad response: the yea-sayer effect (Dinan et al., 2021).

5. Study 2: Retrieval-augmented generation for open-ended Q&A

Having established the safety of our structured conversation in Study 1, we turn our attention to the challenges of open-ended math Q&A. Study 2 explores the use of retrieval-augmented generation to provide accurate and contextually appropriate responses to student queries, building on the safety framework developed in the previous study. We started with the problem of

Table 5: 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???????

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 the first analysis (sec. 5.1), we observed 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 the second analysis (sec. 5.2), we surveyed 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 the third analysis (sec. 5.3), we considered the impact of document relevance on observed preferences.

5.1. Analysis 1: Can We Use Retrieval-augmented Generation and Prompt Engineering to Increase the Groundedness of LLM Responses?

By using retrieval-augmented generation, 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 result in 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 analysis 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.

5.1.1. Guidance Conditions

Prompt engineering is important for LLM performance (Mishra et al., 2022; Lu et al., 2023; Upadhyay et al., 2023). 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" (Yang et al., 2024), 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. 8.2. 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

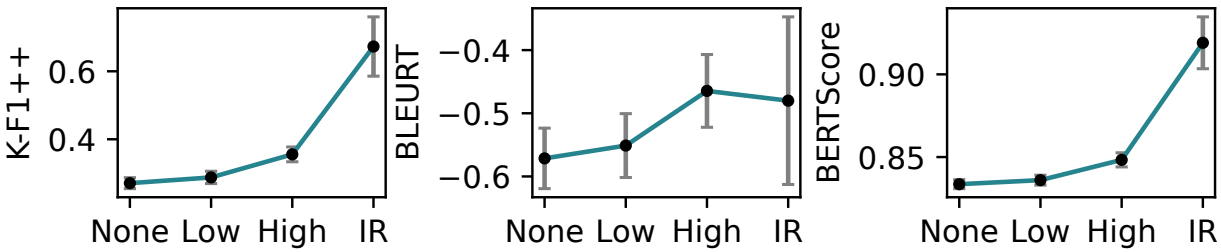


Figure 6: Groundedness for four levels of prompt guidance.

analysis 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."

5.1.2. Student Queries

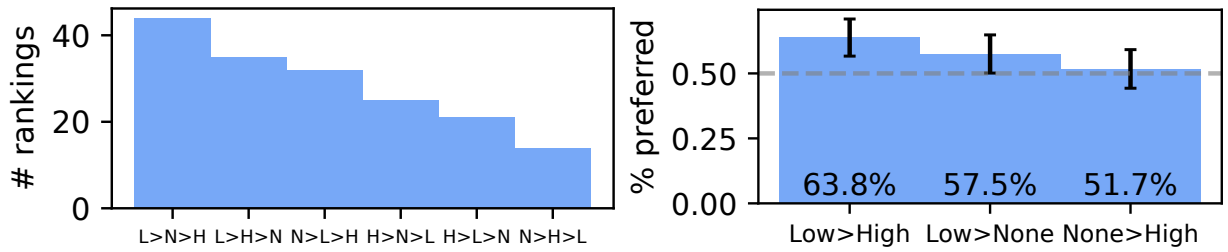
Math Nation is an online math platform with an interactive discussion board (Banawan et al., 2022). On this board, students seek help on math-related questions from their instructors, paid tutors, and peers. We annotated a random sample of 554 Math Nation posts made by students between October 2013 and 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 5.

5.1.3. 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 (Adlakha et al., 2024; Chiesurin et al., 2023; Dziri et al., 2022; Rajpurkar et al., 2016). K-F1++ is a token-level metric that completely ignores semantics, proposed by Chiesurin et al. (2023) as more appropriate for conversational Q&A than Knowledge F1. BERTScore is a token-level metric that uses RoBERTa-base embeddings to model semantics (Zhang et al., 2020). BLEURT is a passage-level metric that models semantics using BERT-base fine-tuned on human relevance judgments (Sellam et al., 2020).

5.1.4. Results

Fig. 6 shows that groundedness 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 Analysis 2, we will directly address the questions of response quality and groundedness by surveying humans.



(a) Distribution of all 144 preference rankings.

(b) Pair-wise preferences from all rankings.

Figure 7: Ranked preferences for LLM responses in three guidance conditions: no guidance (N), low guidance (L), and high guidance (H).

5.2. Analysis 2: Do Humans Prefer More Grounded Responses?

5.2.1. 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 (Adlakha et al., 2024). Responses were spread across four Qualtrics surveys; all questions received 3-4 responses. The survey text is provided in App. 8.3.

5.2.2. Results

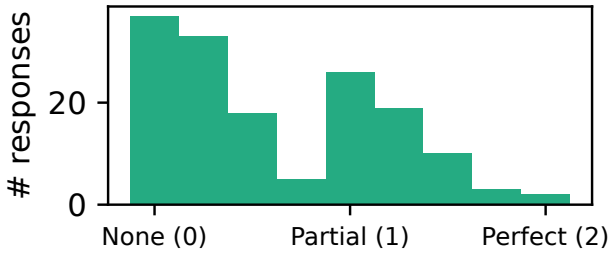
Fig. 7 shows respondent preferences for the three guidance conditions. Responses in the low guidance condition are preferred over responses in the no guidance *and* high guidance conditions. The high and no guidance conditions were statistically indistinguishable. At least two of the guidance conditions 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. 8 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 Analysis 3.[7]

5.3. Analysis 3: How Does Retrieval Relevance Affect Response Groundedness?

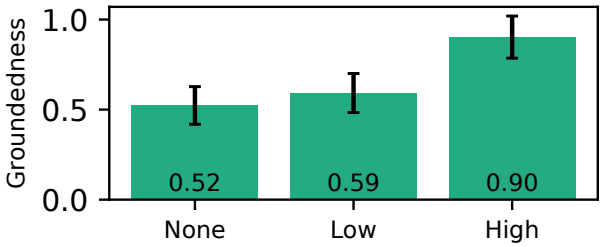
5.3.1. Methods

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

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

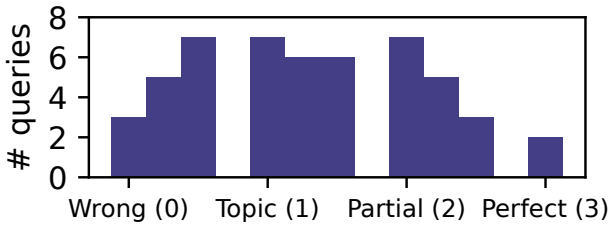


(a) Distribution of groundedness across responses in all three guidance conditions, averaged over annotators.

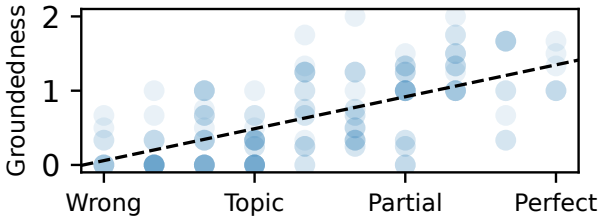


(b) Mean groundedness by guidance condition across all queries and annotators.

Figure 8: Groundedness of the generated responses on an ordinal None (0), Partial (1), Perfect (2) scale.



(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 9: Human-annotated relevance of the retrieved document for all 51 queries.

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 (Hofstätter et al., 2020; Althammer et al., 2022)—see App. 8.4.

5.3.2. 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 retrieved document texts are deemed at least topically relevant, while 33.3% are deemed partially relevant or better; see Fig. 9a for the full distribution. Across all guidance conditions, responses were more likely to be grounded if the retrieved document is relevant (Fig. 9b). 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* (difference in mean relevance=0.19, t=-1.45, p=0.15).

Table 6: 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	Groundedness			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***

5.3.3. Correlation Between Human Annotations and Automated Metrics

Given the results in Analysis 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 6 shows modest positive correlations between automated groundedness metrics and human annotations of relevance. K-F1++ has the strongest correlation (r=0.52) with groundedness, although the correlation is weaker as guidance decreases.

6. Discussion

In this paper, we investigated prompt engineering, a structured conversation framework, and retrieval-augmented generation with LLMs to provide safe and relevant responses for the student users of Intelligent Tutoring Systems (ITSs). We designed a structured conversation and an open-ended Q&A system for an existing math ITS, evaluating their safety and relevance in two studies.

In Study 1, we conducted a student usability test and a field deployment of a structured conversation to teach students about growth mindset. The student usability test was conducted with 109 in-school students across 6 classrooms. We found that the semi-structured conversation design effectively eliminates imposter effects, while our safety filters for student inputs successfully prevent instigator effects (Dinan et al., 2021). We observed no objectionable messages from either the system or the students, and the majority of students rate the conversation positively. During the field deployment with a larger group of students, only 0.31% of student messages receive a concerning moderation score, and most of these are appropriately handled by our system. GPT-3.5 did not generate objectionable outputs, despite sending more than 50,000 messages to students. We found that the design is robust at scale, effectively managing potentially problematic student inputs while maintaining appropriate system responses.

In Study 2, we investigated the use of retrieval-augmented generation (RAG) for answering conceptual math questions. Through a series of analyses involving automated metrics and human evaluations, we find that humans prefer responses when RAG is used, but only if the prompt is not "too guiding." We identified a delicate balance between generating responses preferred by humans and responses closely matched to specific educational resources.

6.1. Study 1: Designing Moderation Systems for Structured Conversations

It was surprisingly straightforward to develop a prompt for GPT-3.5 to respond appropriately to the vast majority of student messages in Study 1 (Narayanan et al., 2023). Instead, our attention was drawn to the more frequent and challenging problem of how to deal with inappropriate or otherwise sensitive student messages. This challenge is analogous to the challenges of content moderation on online platforms, where the context in which a comment exists is important and policies that are reasonable in many cases might be ineffective in edge cases. As an example: how to handle questions regarding contentious political or historical topics? In many cases acknowledging that there are different valid opinions is a good pedagogical approach, but in particularly sensitive or egregious examples this "both-sideism" can be inappropriate (Smith, 2022). However, these are the types of challenges teachers deal with constantly, and we believe that there is a research opportunity here at the intersection of content moderation and classroom management to develop appropriate system actions in response to objectionable student messages.

The specific moderation actions we implemented are reasonable starting points, and by classifying messages at two risk levels we are able to positively redirect conversations with prevetted messages (Leach and Helf, 2016). While these corrective messages were written by educators, in the future we hope that approaches from culturally-responsive classroom management might be combined with soliciting cultural background information from students so that behavioral expectations can be communicated more clearly and correctives can be applied more appropriately (Weinstein et al., 2004; Skiba et al., 2016).

In the event of more serious disclosures, as with the messages we classify as high risk, we argue that our choice to automatically end the conversation and move to human review rather than attempting to generate an appropriate LLM response in the moment is the more ethical one (Stapleton et al., 2024). However, the specific approach we used of ending the conversation is not ideal; we might consider technical infrastructure that starts an in-chat support session with a human or otherwise connects explicitly to contacts at the student's school.

We did observe evidence of the yea-sayer effect in response to some objectionable student messages; future work should explore opportunities for mitigating this effect. In the mean time, designers should monitor for the prevalence of yea-saying and consider technical approaches that explicitly model the appropriate corrective behavior.

Our red-teaming process was effective in its primary goal of identifying potential risk. It had other benefits we did not expect: building organizational confidence. We found that being transparent about the shortcoming of our V1 approach and including designers, educators, and researchers in the evaluation process had the dual benefit of improving trust and soliciting higher-quality feedback to improve the design. We recommend that designers of ITSs conduct similar red-teaming processes.

6.1.1. Limitations & Future Work

The most important limitation of Study 1 is that we evaluated only a single LLM-backed conversation design in a single ITS. The use of a structured conversation may have reduced opportunities for realistic student inputs to trigger inappropriate LLM responses. We hope that future research will explore the potential safety risks of LLMs in ITSs in other contexts and with other LLMs.

The problem of designing for unintended student use is not a new one for ITSs (Baker et al., 2004; Baker, 2007). However, the open-ended inputs and outputs that LLMs enable create new challenges beyond students "gaming the system". It seems likely that there is a long tail of student inputs that LLMs cannot respond to "correctly" 100% of the time (McCoy et al., 2024). Thus, we might take guidance from Baker (2016), keeping ITSs "stupid" and not attempting to respond to all student inputs. Rather than more sophisticated NLP methods for student intent detection, we may instead choose to design highly *structured* ITSs with clear escalation pathways for students who need support that cannot be provided by an LLM.

Understanding potential design trade-offs in the use of LLMs within ITSs requires further research, particularly around moderation actions. It is a limitation of our system that we use a granular, two-level risk assessment for student inputs and provide only a handful of pre-written responses. Future work should explore additional response options and explicitly consider the role humans take in responding to student inputs, perhaps drawing inspiration from work on community moderation (Seering, 2020). Because we observed no serious, high-risk student messages during our field deployment, our work cannot inform best practices on responding to threats of violence or sensitive disclosures.

One limitation of our system design process is that our red-teaming process was somewhat ad-hoc. Similar to Feffer et al. (2024), we found little existing guidance for conducting effective red-teaming exercises. While we believe we identified all relevant potential risks, it is unclear how effective the process we chose will be for other LLM-related use cases, especially given the low baseline rate of inappropriate responses we observed from GPT-3.5. Future work should adapt red-teaming more explicitly to multi-stakeholder education environments involving students, parents, and educators.

6.2. Study 2: Groundedness Preferences for Conceptual Q&A

In Study 2, we found that RAG *can* improve response quality, but too much groundedness can decrease quality. We argue that designers of math Q&A systems should consider trade-offs between generating responses preferred by humans and responses closely matched to specific educational resources. Math Q&A 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. (2023) 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. Faithfulness is a desirable property for conceptual math Q&A 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., see Adlakha et al. 2024; Dziri et al. 2022), and explaining the role of retrieved document relevance (as in our surprising Analysis 3 results finding that relevance was not a meaningful predictor of human preference).

6.2.1. Limitations & Future Work

Study 2 is a preliminary step toward understanding the relationship between groundedness and preference in conceptual math Q&A systems. Future work must extend beyond single-turn responses to include exploration of follow-up questions (Wang et al., 2024) and to design for the

actual context of use. The most important limitation of Study 2 is that we did not collect preferences directly from students, although we did use real student questions. Future qualitative research of students' preferences should focus not only on correctness but also on factors such as conceptual granularity, curricular alignment, and cultural relevance. Beyond preferences, future math Q&A systems that use RAG will need to explore the relationship between students' response preferences and actual learning outcomes. We are introducing Q&A alongside structured math practice in Rori, using the continual safety processes we designed to ensure that Q&A is useful for students.

7. Conclusion

We conducted a field study with more than 8,000 students using a conversational LLM to learn about growth mindset, finding that the GPT-3.5 LLM produced few inappropriate messages. A structured conversation design process for a straightforward moderation system was additionally effective at detecting inappropriate student messages. A second study demonstrated that retrieval-augmented generation was effective at improving a conceptual Q&A system using GPT-3.5. Our two studies together suggest a focus on future work evaluating the effectiveness of LLMs in real ITSs. While this work was preliminary in many ways, we argue that a combination of iterative red-teaming, prompt engineering, and moderation system design are sufficient for ethically deploying LLM-backed research prototypes in ITSs. As additional LLM prototypes are incorporated into the design of ITSs, we should focus on evaluating whether LLM-generated responses to students align with the classroom management goals of the education system using the ITS. More effective evaluation processes will enable the rigorous evaluation of LLM suitability for particular ITS tasks, from open-ended Q&A to focused teaching of a single concept.

Acknowledgements

We would like to thank Ralph Abboud, Nessie Kozhakhmetova, Bill Roberts, Hannah Horne-Robinson, 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 by the Digital Harbor Foundation.

References

Adlakha, V., BehnamGhader, P., Lu, X. H., Meade, N., and Reddy, S. 2024. Evaluating Correctness and Faithfulness of Instruction-Following Models for Question Answering. *Transactions of the Association for Computational Linguistics 12*, 681–699. Cambridge, MA. MIT Press.

Agiza, A., Mostagir, M., and Reda, S. 2024. PoliTune: Analyzing the Impact of Data Selection and Fine-Tuning on Economic and Political Biases in Large Language Models. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 7*, 2–12.

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. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, N. Wang, G. Rebolledo-Mendez,

V. Dimitrova, N. Matsuda, and O. C. Santos, Eds. Communications in Computer and Information Science. Springer Nature Switzerland, Cham, 26–31.

ALEVEN, V. AND KOEDINGER, K. R. 2001. Investigations into Help Seeking and Learning with a Cognitive Tutor. *Working Notes of the AIED 2001 Workhop "Help Provision And Help Seeking In Interactive Learning Environments"*.

ALTHAMMER, S., HOFSTÄTTER, S., VERBERNE, S., AND HANBURY, A. 2022. TripJudge: A Relevance Judgement Test Collection for TripClick Health Retrieval. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, M. A. Hasan and L. Xiong, Eds. ACM, Atlanta GA USA, 3801–3805.

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, 135–152. IOS Press.

BAKER, R. S. 2007. Modeling and understanding students' off-task behavior in intelligent tutoring systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, M. B. Rosson and D. Gilmore, Eds. CHI '07. Association for Computing Machinery, New York, NY, USA, 1059–1068.

BAKER, R. S. 2016. Stupid Tutoring Systems, Intelligent Humans. *International Journal of Artificial Intelligence in Education 26,* 2 (June), 600–614.

BAKER, R. S., CORBETT, A. T., AND KOEDINGER, K. R. 2004. Detecting Student Misuse of Intelligent Tutoring Systems. In *Intelligent Tutoring Systems*, J. C. Lester, R. M. Vicari, and F. Paraguaçu, Eds. Springer, Berlin, Heidelberg, 531–540.

BAKER, R. S. AND HAWN, A. 2022. Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education 32,* 4 (Dec.), 1052–1092.

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). In *Proceedings of the Ninth ACM Conference on Learning @ Scale*, R. F. Kizilcec, K. Davis, and X. Ochoa, Eds. L@S '22. Association for Computing Machinery, New York, NY, USA, 389–394.

BASTANI, H., BASTANI, O., SUNGU, A., GE, H., KABAKCI, O., AND MARIMAN, R. 2024. Generative AI Can Harm Learning. Available at SSRN.

BECK, K. A., OGLOFF, J. R. P., AND CORBISHLEY, A. 1994. Knowledge, Compliance, and Attitudes of Teachers toward Mandatory Child Abuse Reporting in British Columbia. *Canadian Journal of Education / Revue canadienne de l'éducation 19,* 1, 15–29. Canadian Society for the Study of Education.

BERGER, E., CHIONH, N., AND MIKO, A. 2022. School Leaders' Experiences on Dealing with Students Exposed to Domestic Violence. *Journal of Family Violence 37,* 7 (Oct.), 1089–1100.

BORGEAUD, S., MENSCH, A., HOFFMANN, J., CAI, T., RUTHERFORD, E., MILLICAN, K., VAN DEN DRIESSCHE, G. B., LESPIAU, J.-B., DAMOC, B., CLARK, A., DE LAS CASAS, D., GUY, A., MENICK, J., RING, R., HENNIGAN, T., HUANG, S., MAGGIORE, L., JONES, C., CASSIRER, A., BROCK, A., PAGANINI, M., IRVING, G., VINYALS, O., OSINDERO, S., SIMONYAN, K., RAE, J., ELSEN, E., AND SIFRE, L. 2022. Improving Language Models by Retrieving from Trillions of Tokens. In *Proceedings of the 39th International Conference on Machine Learning*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds. Proceedings of Machine Learning Research, vol. 162. PMLR, 2206–2240.

CAINES, A., BENEDETTO, L., TASLIMIPOOR, S., DAVIS, C., GAO, Y., ANDERSEN, O. E., YUAN, Z., ELLIOTT, M., MOORE, R., BRYANT, C., REI, M., YANNAKOUDAKIS, H., MULLOOLY, A.,

NICHOLLS, D., AND BUTTERY, P. 2023. On the Application of Large Language Models for Language Teaching and Assessment Technology. In *LLM@AIED*, S. Moore, J. Stamper, R. Tong, C. Cao, Z. Liu, X. Hu, Y. Lu, J. Liang, H. Khosravi, P. Denny, A. Singh, and C. Brooks, Eds. 173–197.

CHASE, H. 2023. How to use the Parent Document Retriever. https://python.langchain.com/docs/how_to/parent_document_retriever/.

CHEN, D., FISCH, A., WESTON, J., AND BORDES, A. 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)*, R. Barzilay and M.-Y. Kan, Eds. Association for Computational Linguistics, Vancouver, Canada, 1870–1879.

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. In *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Association for Computational Linguistics, Toronto, Canada, 947–959.

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, 105–120.

DEMSZKY, D., LIU, J., HILL, H. C., JURAFSKY, D., AND PIECH, C. 2024. Can Automated Feedback Improve Teachers' Uptake of Student Ideas? Evidence From a Randomized Controlled Trial in a Large-Scale Online Course. *Educational Evaluation and Policy Analysis 46,* 3 (Sept.), 483–505. American Educational Research Association.

DINAN, E., ABERCROMBIE, G., BERGMAN, A. S., SPRUIT, S., HOVY, D., BOUREAU, Y.-L., AND RIESER, V. 2021. Anticipating Safety Issues in E2E Conversational AI: Framework and Tooling. arXiv:2107.03451 [cs].

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. *Transactions of the Association for Computational Linguistics 10*, 1473–1490. MIT Press.

EMMER, E. T. AND STOUGH, L. M. 2001. Classroom Management: A Critical Part of Educational Psychology, With Implications for Teacher Education. *Educational Psychologist 36,* 2 (June), 103–112. Routledge.

FALKINER, M., THOMSON, D., AND DAY, A. 2017. Teachers' Understanding and Practice of Mandatory Reporting of Child Maltreatment. *Children Australia 42,* 1 (Mar.), 38–48.

FALKINER, M., THOMSON, D., GUADAGNO, B., AND DAY, A. 2020. Heads you win, tails I lose: The dilemma mandatory reporting poses for teachers. *Australian Journal of Teacher Education (Online) 42,* 9 (Aug.), 93–110. Edith Cowan University.

FEFFER, M., SINHA, A., DENG, W. H., LIPTON, Z. C., AND HEIDARI, H. 2024. Red-Teaming for Generative AI: Silver Bullet or Security Theater? *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 7*, 421–437.

FISCHER, J. E. 2023. Generative AI Considered Harmful. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, M. Lee, C. Munteanu, M. Porcheron, J. Trippas, and S. T. Völkel, Eds. CUI '23. Association for Computing Machinery, New York, NY, USA, 1–5.

GABRIEL, I., MANZINI, A., KEELING, G., HENDRICKS, L. A., RIESER, V., IQBAL, H., TOMAŠEV, N., KTENA, I., KENTON, Z., RODRIGUEZ, M., EL-SAYED, S., BROWN, S., AKBULUT, C., TRASK, A., HUGHES, E., BERGMAN, A. S., SHELBY, R., MARCHAL, N., GRIFFIN, C., MATEOS-GARCIA, J., WEIDINGER, L., STREET, W., LANGE, B., INGERMAN, A., LENTZ, A., ENGER, R., BARAKAT,

A., KRAKOVNA, V., SIY, J. O., KURTH-NELSON, Z., MCCROSKERY, A., BOLINA, V., LAW, H., SHANAHAN, M., ALBERTS, L., BALLE, B., HAAS, S. D., IBITOYE, Y., DAFOE, A., GOLDBERG, B., KRIER, S., REESE, A., WITHERSPOON, S., HAWKINS, W., RAUH, M., WALLACE, D., FRANKLIN, M., GOLDSTEIN, J. A., LEHMAN, J., KLENK, M., VALLOR, S., BILES, C., MORRIS, M. R., KING, H., ARCAS, B. A. Y., ISAAC, W., AND MANYIKA, J. 2024. The Ethics of Advanced AI Assistants. arXiv:2404.16244.

GIRARD, S. AND JOHNSON, H. 2010. What Do Children Favor as Embodied Pedagogical Agents? In *Intelligent Tutoring Systems*, V. Aleven, J. Kay, and J. Mostow, Eds. Springer, Berlin, Heidelberg, 307–316.

GOLDMAN, E. J., BAUMANN, A.-E., AND POULIN-DUBOIS, D. 2023. Preschoolers' anthropomorphizing of robots: Do human-like properties matter? *Frontiers in Psychology 13*.

GOLDMAN, J. D. G. 2007. Primary school student-teachers' knowledge and understandings of child sexual abuse and its mandatory reporting. *International Journal of Educational Research 46,* 6 (Jan.), 368–381.

GOOD, T. L. AND BROPHY, J. E. 2008. *Looking in Classrooms*. Pearson/Allyn and Bacon.

GRAESSER, A. C., PERSON, N. K., AND MAGLIANO, J. P. 1995. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology 9,* 6, 495–522.

GREENE, R., SANDERS, T., WENG, L., AND NEELAKANTAN, A. 2022. New and improved embedding model. https://openai.com/blog/new-and-improved-embedding-model.

GUU, K., LEE, K., TUNG, Z., PASUPAT, P., AND CHANG, M.-W. 2020. REALM: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, H. Daumé and A. Singh, Eds. ICML'20, vol. 119. JMLR.org, 3929–3938.

HANEY, M. R. 2004. Ethical Dilemmas Associated With Self-Disclosure in Student Writing. *Teaching of Psychology*. 31(3), 167–171.

HENDRYCKS, D., CARLINI, N., SCHULMAN, J., AND STEINHARDT, J. 2022. Unsolved Problems in ML Safety. arXiv:2109.13916.

HENKEL, O., HILLS, L., BOXER, A., ROBERTS, B., AND LEVONIAN, Z. 2024. Can Large Language Models Make the Grade? An Empirical Study Evaluating LLMs Ability To Mark Short Answer Questions in K-12 Education. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, D. Joyner, M. K. Kim, X. Wang, and M. Xia, Eds. L@S '24. Association for Computing Machinery, New York, NY, USA, 300–304.

HENKEL, O., HORNE-ROBINSON, H., KOZHAKHMETOVA, N., AND LEE, A. 2024. Effective and Scalable Math Support: Evidence on the Impact of an AI- Tutor on Math Achievement in Ghana. arXiv:2402.09809 [cs].

HENKEL, O., LEVONIAN, Z., LI, C., AND POSTLE, M. 2024. Retrieval-augmented Generation to Improve Math Question-Answering: Trade-offs Between Groundedness and Human Preference. In *Proceedings of the 17th International Conference on Educational Data Mining*, C. D. Epp, B. Paassen, and D. Joyner, Eds. International Educational Data Mining Society, Atlanta, GA, USA, 315–320.

HOBERT, S. AND WOLFF, R. M. V. 2019. Say Hello to Your New Automated Tutor – A Structured Literature Review on Pedagogical Conversational Agents. *Wirtschaftsinformatik 2019 Proceedings*.

HOFSTÄTTER, S., ZLABINGER, M., SERTKAN, M., SCHRÖDER, M., AND HANBURY, A. 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*, M. d'Aquin, S. Dietze, C. Hauff, E. Curry, and P. C. Mauroux, Eds. CIKM '20. Association for Computing Machinery, New York, NY, USA, 3031–3038.

HOLSTEIN, K., MCLAREN, B. M., AND ALEVEN, V. 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*, A. Wise, P. H. Winne, G. Lynch, X. Ochoa, I. Molenaar, S. Dawson, and M. Hatala, Eds. LAK '17. Association for Computing Machinery, New York, NY, USA, 257–266.

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, 57–71.

IZACARD, G., LEWIS, P., LOMELI, M., HOSSEINI, L., PETRONI, F., SCHICK, T., DWIVEDI-YU, J., JOULIN, A., RIEDEL, S., AND GRAVE, E. 2024. Atlas: few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res. 24,* 1 (Mar.), 251:11912–251:11954.

JHAVER, S., ZHANG, A. Q., CHEN, Q. Z., NATARAJAN, N., WANG, R., AND ZHANG, A. X. 2023. Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor. *Proceedings of the ACM on Human-Computer Interaction 7,* CSCW2 (Oct.), 289:1–289:33.

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, 1–38.

KARUMBAIAH, S., LIZARRALDE, R., ALLESSIO, D., WOOLF, B., ARROYO, I., AND WIXON, N. 2017. Addressing Student Behavior and Affect with Empathy and Growth Mindset. In *Proceedings of the International Conference on Educational Data Mining*, Hu, Xiangen, Barnes, Tiffany, Hershkovitz, Arnon, and Paquette, Luc, Eds. International Educational Data Mining Society, Wuhan, China, 96–103.

KASNECI, E., SESSLER, K., KÜCHEMANN, S., BANNERT, M., DEMENTIEVA, D., FISCHER, F., GASSER, U., GROH, G., GÜNNEMANN, S., HÜLLERMEIER, E., KRUSCHE, S., KUTYNIOK, G., MICHAELI, T., NERDEL, C., PFEFFER, J., POQUET, O., SAILER, M., SCHMIDT, A., SEIDEL, T., STADLER, M., WELLER, J., KUHN, J., AND KASNECI, G. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences 103*, 102274.

KHOSRAWI-RAD, B., RINN, H., SCHLIMBACH, R., GEBBING, P., YANG, X., LATTEMANN, C., MARKGRAF, D., AND ROBRA-BISSANTZ, S. 2022. Conversational Agents in Education – A Systematic Literature Review. *ECIS 2022 Research Papers*. Article 18.

KIZILCEC, R. F. AND GOLDFARB, D. 2019. Growth Mindset Predicts Student Achievement and Behavior in Mobile Learning. In *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale*. L@S '19. Association for Computing Machinery, New York, NY, USA, 1–10.

KRAFT, M. A. AND FALKEN, G. T. 2021. A blueprint for scaling tutoring and mentoring across public schools. *AERA Open 7,* 1, 1–21. Los Angeles, CA. SAGE Publications.

LAZARIDOU, A., GRIBOVSKAYA, E., STOKOWIEC, W. J., AND GRIGOREV, N. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering. arXiv:2203.05115 [cs.CL].

LEACH, D. AND HELF, S. 2016. Using a Hierarchy of Supportive Consequences to Address Problem Behaviors in the Classroom. *Intervention in School and Clinic 52,* 1 (Sept.), 29–33. SAGE Publications.

LEHMANN, M., CORNELIUS, P. B., AND STING, F. J. 2024. AI Meets the Classroom: When Does ChatGPT Harm Learning? arXiv:2409.09047.

LEVIN, J. AND NOLAN, J. F. 2002. *What Every Teacher Should Know About Classroom Management*, 1st ed. Pearson.

LEVONIAN, Z. AND HENKEL, O. 2024. Safe Generative Chats in a WhatsApp Intelligent Tutoring System. In *Joint Proceedings of the Human-Centric eXplainable AI in Education and the Leveraging Large Language Models for Next Generation Educational Technologies Workshops (HEXED-L3MNGET 2024) co-located with 17th International Conference on Educational Data Mining (EDM 2024)*, J. D. Pinto, E. Worden, A. Botelho, L. Cohausz, C. Cohn, M. Feng, N. Heffernan, A. Hellas, L. Jiang, D. Joyner, T. Käser, J. Kim, A. Lan, C. Li, J. Littenberg-Tobias, Q. Liu, C. MacLellan, S. Moore, M. Pankiewicz, L. Paquette, Z. A. Pardos, A. Rafferty, A. Singla, S. Sonkar, V. Swamy, R. E. Wang, and C. Walkington, Eds. Atlanta, GA, USA. arXiv:2407.04915 [cs].

LEVONIAN, Z., HENKEL, O., AND ROBERTS, B. 2023. llm-math-education: Retrieval augmented generation for middle-school math question answering and hint generation. https://zenodo.org/record/8284412.

LEWIS, P., PEREZ, E., PIKTUS, A., PETRONI, F., KARPUKHIN, V., GOYAL, N., KÜTTLER, H., LEWIS, M., YIH, W.-T., ROCKTÄSCHEL, T., RIEDEL, S., AND KIELA, D. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds. NeurIPS'20. Curran Associates Inc., Red Hook, NY, USA, 9459–9474.

LI, C. AND XING, W. 2021. Natural language generation using deep learning to support MOOC learners. *International Journal of Artificial Intelligence in Education 31*, 186–214. Springer.

LIANG, P. P., WU, C., MORENCY, L.-P., AND SALAKHUTDINOV, R. 2021. Towards Understanding and Mitigating Social Biases in Language Models. In *Proceedings of the 38th International Conference on Machine Learning*, Marina Meila and Tong Zhang, Eds. PMLR, 6565–6576. ISSN: 2640-3498.

LIEB, A. AND GOEL, T. 2024. Student Interaction with NewtBot: An LLM-as-tutor Chatbot for Secondary Physics Education. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, F. F. Mueller, P. Kyburz, J. R. Williamson, and C. Sas, Eds. CHI EA '24. Association for Computing Machinery, New York, NY, USA, 1–8.

LIN, J., MA, X., LIN, S.-C., YANG, J.-H., PRADEEP, R., AND NOGUEIRA, R. 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*, F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, and T. Sakai, Eds. SIGIR '21. Association for Computing Machinery, New York, NY, USA, 2356–2362.

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:2309.01809 [cs].

MAITI, P. AND GOEL, A. K. 2024. How Do Students Interact with an LLM-powered Virtual Teaching Assistant in Different Educational Settings? In *Proceedings of the Seventeenth International Conference on Educational Data Mining (EDM) Workshop: Leveraging LLMs for Next Generation Educational Technologies*, J. D. Pinto, E. Worden, A. Botelho, L. Cohausz, C. Cohn, M. Feng, N. Heffernan, A. Hellas, L. Jiang, D. Joyner, T. Käser, J. Kim, A. Lan, C. Li, J. Littenberg-Tobias, Q. Liu, C. MacLellan, S. Moore, M. Pankiewicz, L. Paquette, Z. A. Pardos, A. Rafferty, A. Singla, S. Sonkar, V. Swamy, R. E. Wang, and C. Walkington, Eds. educationaldatamining.org.

MARECEK, L., ANTHONY-SMITH, M., AND HONEYCUTT MATHIS, A. 2020. *Prealgebra*, 2 ed. OpenStax.

MARZANO, R. J. 2005. *A Handbook for Classroom Management that Works*. ASCD. Google-Books-ID: BMOQFLa0fcEC.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., and Griffiths, T. L. 2024. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences 121,* 41 (Oct.), e2322420121.

Mialon, G., Dessi, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Roziere, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., Grave, E., LeCun, Y., and Scialom, T. 2023. Augmented Language Models: a Survey. *Transactions on Machine Learning Research.*

Mishra, S., Khashabi, D., Baral, C., Choi, Y., and Hajishirzi, H. 2022. Reframing Instructional Prompts to GPTk's Language. In *Findings of the Association for Computational Linguistics: ACL 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Association for Computational Linguistics, Dublin, Ireland, 589–612.

Moschkovich, J. N. 2015. Scaffolding student participation in mathematical practices. *Zdm 47*, 1067–1078. Springer.

Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. 2023. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, Eds. Association for Computational Linguistics, Dubrovnik, Croatia, 2014–2037.

Murphy, P. K. and Alexander, P. A. 2013. Situating Text, Talk, and Transfer in Conceptual Change: Concluding Thoughts. In *International Handbook of Research on Conceptual Change*, 2 ed. Routledge, 603–621.

NAEP. 2022. NAEP Mathematics: National Average Scores.

Narayanan, A., Kapoor, S., and Lazar, S. 2023. Model alignment protects against accidental harms, not intentional ones. https://www.aisnakeoil.com/p/model-alignment-protects-against.

Navigli, R., Conia, S., and Ross, B. 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. *Journal of Data and Information Quality 15,* 2 (June), 10:1–10:21.

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. https://www.nber.org/papers/w27476.

Nie, A., Chandak, Y., Suzara, M., Ali, M., Woodrow, J., Peng, M., Sahami, M., Brunskill, E., and Piech, C. 2024. The GPT Surprise: Offering Large Language Model Chat in a Massive Coding Class Reduced Engagement but Increased Adopters Exam Performances. arXiv:2407.09975.

Nye, B. D., Mee, D., and Core, M. G. 2023. Generative Large Language Models for Dialog-Based Tutoring: An Early Consideration of Opportunities and Concerns. In *Proceedings of the Workshop on Empowering Education with LLMs - the Next-Gen Interface and Content Generation 2023 co-located with 24th International Conference on Artificial Intelligence in Education (AIED 2023)*, S. Moore, J. Stamper, R. Tong, C. Cao, Z. Liu, X. Hu, Y. Lu, J. Liang, H. Khosravi, P. Denny, A. Singh, and C. Brooks, Eds. Tokyo, Japan.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs].

OpenAI. 2024. Moderation: OpenAI API. https://platform.openai.com/docs/guides/moderation/overview.

Pardos, Z. A. and Bhandari, S. 2023. Learning gain differences between chatgpt and human tutor generated algebra hints. *CoRR abs/2302.06871*.

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:2302.12813 [cs].

PSOTKA, J., MASSEY, L. D., AND MUTTER, S. A., Eds. 1988. *Intelligent tutoring systems: Lessons learned*. Lawrence Erlbaum Associates, Hillsdale, NJ, US.

RAJPURKAR, P., ZHANG, J., LOPYREV, K., AND LIANG, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, J. Su, K. Duh, and X. Carreras, Eds. Association for Computational Linguistics, Austin, Texas, 2383–2392.

RITTER, S., ANDERSON, J. R., KOEDINGER, K. R., AND CORBETT, A. 2007. Cognitive Tutor: Applied research in mathematics education. *Psychonomic bulletin & review 14*, 249–255. Springer.

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*, 587–597. Springer.

RODRIGO, M. M. T., BAKER, R. S. J. D., AND ROSSI, L. 2013. Student Off-Task Behavior in Computer-Based Learning in the Philippines: Comparison to Prior Research in the USA. *Teachers College Record 115,* 10 (Oct.), 1–27. SAGE Publications.

SABORNIE, E. J. AND ESPELAGE, D. L. 2022. *Handbook of Classroom Management*, 3rd ed. Routledge.

SEDLMEIER, P. 2001. Intelligent Tutoring Systems. In *International Encyclopedia of the Social & Behavioral Sciences*, N. J. Smelser and P. B. Baltes, Eds. Pergamon, Oxford, 7674–7678.

SEERING, J. 2020. Reconsidering Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proc. ACM Hum.-Comput. Interact. 4,* CSCW2 (Oct.), 107:1–107:28.

SELLAM, T., DAS, D., AND PARIKH, A. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Association for Computational Linguistics, Online, 7881–7892.

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. In *Proceedings of the NeurIPS 2021 Math AI for Education Workshop*, Pan Lu, Yuhuai Wu, Sean Welleck, Xiaodan Liang, Eric Xing, and James McClelland, Eds. neurips.cc, Virtual, 1–11.

SKIBA, R., ORMISTON, H., MARTINEZ, S., AND CUMMINGS, J. 2016. Teaching the Social Curriculum: Classroom Management as Behavioral Instruction. *Theory Into Practice 55,* 2 (Apr.), 120–128.

SMITH, R. 2022. How "both-sideism" harms health. *BMJ 378*, o2136. British Medical Journal Publishing Group, Opinion.

SONKAR, S., LIU, N., MALLICK, D., AND BARANIUK, R. 2023. CLASS: A Design Framework for Building Intelligent Tutoring Systems Based on Learning Science principles. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Association for Computational Linguistics, Singapore, 1941–1961.

SOTTILARE, R. A., GRAESSER, A., HU, X., AND GOLDBERG, B. S. 2014. Design Recommendations for Intelligent Tutoring Systems. Volume 2: Instructional Management. Tech. rep., University of Southern California Los Angeles. Jan. Technical Reports.

STAPLETON, L., LIU, S., LIU, C., HONG, I., CHANCELLOR, S., KRAUT, R. E., AND ZHU, H. 2024. "If This Person is Suicidal, What Do I Do?": Designing Computational Approaches to Help Online Volunteers Respond to Suicidality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, F. F. Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. T. Dugas, and I. Shklovski, Eds. CHI '24. Association for Computing Machinery, New York, NY, USA. Honolulu, HI, USA.

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*, 210–226. Springer.

TAO, Y., VIBERG, O., BAKER, R. S., AND KIZILCEC, R. F. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus 3,* 9 (Sept.), pgae346.

UPADHYAY, S., GINSBERG, E., AND CALLISON-BURCH, C. 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)*, E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, and T. Zesch, Eds. Association for Computational Linguistics, Toronto, Canada, 20–28.

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

VANLEHN, K. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist 46,* 4, 197–221. Taylor & Francis.

WANG, R. E., RIBEIRO, A. T., ROBINSON, C. D., LOEB, S., AND DEMSZKY, D. 2024. Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise. arXiv:2410.03017.

WANG, X., WANG, Z., LIU, J., CHEN, Y., YUAN, L., PENG, H., AND JI, H. 2024. MINT: Evaluating LLMs in Multi-turn Interaction with Tools and Language Feedback. In *The Twelfth International Conference on Learning Representations*, B. Kim, S. Chaudhuri, K. Fragkiadaki, M. E. Khan, and Y. Sun, Eds.

WEINSTEIN, C. S., TOMLINSON-CLARKE, S., AND CURRAN, M. 2004. Toward a Conception of Culturally Responsive Classroom Management. *Journal of Teacher Education 55,* 1 (Jan.), 25–38.

XU, Y., HU, L., ZHAO, J., QIU, Z., YE, Y., AND GU, H. 2024. A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias. arXiv:2404.00929 [cs].

YAN, L., SHA, L., ZHAO, L., LI, Y., MARTINEZ-MALDONADO, R., CHEN, G., LI, X., JIN, Y., AND GAŠEVIĆ, D. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology 55,* 1, 90–112.

YANG, C., WANG, X., LU, Y., LIU, H., LE, Q. V., ZHOU, D., AND CHEN, X. 2024. Large Language Models as Optimizers. In *The Twelfth International Conference on Learning Representations*, B. Kim, S. Chaudhuri, K. Fragkiadaki, M. E. Khan, and Y. Sun, Eds.

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. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, E. Denton, J.-W. Ha, and J. Vanschoren, Eds.

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.), 119:1–119:24.

YANG, W., LU, K., YANG, P., AND LIN, J. 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*, B. Piwowarski, M. Chevalier, E. Gaussier, Y. Maarek, J.-Y. Nie, and F. Scholer, Eds. SIGIR'19. Association for Computing Machinery, New York, NY, USA, 1129–1132.

YEAGER, D. S., ROMERO, C., PAUNESKU, D., HULLEMAN, C. S., SCHNEIDER, B., HINOJOSA, C., LEE, H. Y., O'BRIEN, J., FLINT, K., ROBERTS, A., TROTT, J., GREENE, D., WALTON, G. M., AND DWECK, C. S. 2016. Using Design Thinking to Improve Psychological Interventions: The Case of the Growth Mindset During the Transition to High School. *Journal of Educational Psychology 108,* 3 (Apr.), 374–391.

ZAMANI, H., DIAZ, F., DEHGHANI, M., METZLER, D., AND BENDERSKY, M. 2022. Retrieval-Enhanced Machine Learning. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, E. Amigo, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, and G. Kazai, Eds. SIGIR '22. Association for Computing Machinery, New York, NY, USA, 2875–2886.

ZHANG, T., KISHORE, V., WU, F., WEINBERGER, K. Q., AND ARTZI, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the International Conference on Learning Representations*, A. Rush, S. Mohamed, D. Song, K. Cho, and M. White, Eds. Virtual.

8. APPENDICES

8.1. 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. GPT-4o and GPT-4o-mini were not released at the time of the study. 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 (Lin et al., 2021; Yang et al., 2019).

8.2. 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.

8.2.1. 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.

8.2.2. 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:
===
{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.

8.2.3. 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.

8.2.4. 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}

8.3. 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 7 shows per-annotator counts.

8.3.1. Intro Page

This survey will consist of 15 questions. Your progress will save after each question.
Who are you? (Annotator name) _____

8.3.2. Query Page

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

Table 7: 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

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}	◯	◯	◯

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	◯	◯	◯

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

8.4. RELEVANCE SURVEY

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

8.4.1. 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.

8.4.2. Spreadsheet Tab

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