Step-by-Step Remediation of Students' Mathematical Mistakes

Rose E. Wang Qingyang Zhang Carly Robinson
Susanna Loeb Dorottya Demszky
Stanford University
{rewang,ddemszky}@stanford.edu

Abstract

Scaling high-quality tutoring is a major challenge in education. Because of the growing demand, many platforms employ novice tutors who, unlike professional educators, struggle to effectively address student mistakes and thus fail to seize prime learning opportunities for students. In this paper, we explore the potential for large language models (LLMs) to assist math tutors in remediating student mistakes. We present `ReMath`, a benchmark co-developed with experienced math teachers that deconstructs their thought process for remediation. The benchmark consists of three step-by-step tasks: (1) infer the type of student error, (2) determine the strategy to address the error, and (3) generate a response that incorporates that information. We evaluate the performance of state-of-the-art instruct-tuned and dialog models on `ReMath`. Our findings suggest that although models consistently improve upon original tutor responses, we cannot rely on models alone to remediate mistakes. Providing models with the error type (e.g., the student is guessing) and strategy (e.g., simplify the problem) leads to a 75% improvement in the response quality over models without that information. Nonetheless, despite the improvement, the quality of the best model's responses still falls short of experienced math teachers. Our work sheds light on the potential and current limitations of using LLMs to provide high-quality learning experiences for both tutors and students at scale.

1 Introduction

If you can both listen to children and accept their answers not as things to just be judged right or wrong but as pieces of information which may reveal what the child is thinking, you will have taken a giant step toward becoming a master teacher rather than merely a disseminator of information.

— J. A. EASLEY, JR. & R, E. ZWOYER (1975)

Human tutoring plays a critical role in accelerating student learning, and is one of the primary ways to combat recent pandemic-related learning losses (Fryer Jr and Howard-Noveck, 2020; Nickow et al., 2020; Robinson and Loeb, 2021; of Education, 2021; Accelerator, 2022). To accommodate the growing demand for tutoring, many tutoring providers engage novice tutors. While novice tutors may exercise the domain knowledge, they often lack the specialized training of professional educators in interacting with students. However, research suggests that, with proper training, many people can serve as effective tutors—not just trained educators (Nickow et al., 2020).

Responding to student mistakes in real-time is a critical area of tutoring where novice tutors tend to struggle. Mistakes are prime learning opportunities to address misconceptions (Boaler, 2013), but responding to them effectively requires expertise in pedagogical techniques to probe students' thinking (Shaughnessy et al., 2021). The importance of fostering positive educator-student relationships amplify this challenge. Prior research has shown that positive educator-student relationships improve student outcomes (Pianta, 2016; Roorda et al., 2011). The way that educators respond to student errors can shape how students perceive themselves, which subsequently impacts their engagement in the learning process (Robinson, 2022). Therefore, effective tutors not only provide *useful* responses, but also *caring* responses to remediate student mistakes.

Experienced educators could provide high-quality feedback to novice tutors. However, hiring experienced educators to provide timely feedback is resource-intensive (Kraft et al., 2018; Kelly et al., 2020). A potential solution is the use of automated tutors (Graesser et al., 2004). With recent advances in large language models (LLMs), such as ChatGPT or GPT4, this approach has gained even more interest (Khan Academy, 2023). However, even though LLMs have the potential to address the *scal-*

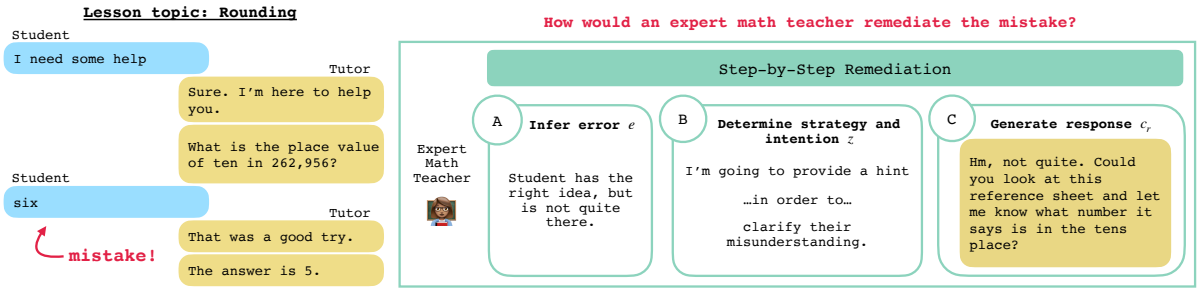


Figure 1: Illustration of the ReMath benchmark. The left handside shows the original tutor-student conversation on the lesson topic, rounding. The last message is the student's mistake that should be remediated. The right handside shows the original tutor's response and the math teacher's step-by-step remediation annotations. The benchmark tasks are Task A: the inferred error that the student makes (e), Task B: the strategy and desired outcome of the strategy (z), and Task C: the remediation response (c_r).

ability challenge faced by tutoring platforms, their ability to effectively remediate student mistakes is yet to be evaluated. Prior work suggests that they have several shortcomings, including lacking reliable subject knowledge (Frieder et al., 2023) and pedagogical expertise (Wang and Demszky, 2023). They also suffer from hallucination, generating text that is factually incorrect (Ji et al., 2023).

To address these challenges, our work makes the following key contributions:

1. We collaborate closely with experienced teachers to develop the ReMath framework. ReMath breaks down the process of remediating student mistakes into three tasks: **Task A**: Infer the type of student error, **Task B**: Determine the strategy and intention behind the strategy for remediation, and **Task C**: Generate a response to the student's mistake.

2. We contribute a dataset of original tutor responses and expert teacher annotations using ReMath. Our dataset consists of tutor-student interactions from chat-based tutoring sessions conducted with 1st through 5th grade students in Title I schools that predominantly serve low-income students of color. Each example is annotated by two expert teachers.

3. Using our framework and dataset, we conduct a thorough evaluation, combining automated and human approaches, of LLMs and teacher responses to student mistakes. To our knowledge, our work is the first to assess the performance of LLMs—such as GPT4 and open-sourced models—on remediating student mistakes. We find that models' responses improve significantly over the original tutors' responses but they still fall short of responses written by experienced teachers. Providing the model with the error type and strategy improves response quality, close to that of the expert teachers. This suggests that a promising approach lies in combining the strengths of LLMs models with teacher expertise to scale and improve tutoring effectiveness.

2 Related Work

2.1 Responding to Student Mistakes in Mathematics

The ability of teachers to identify student errors is essential for effectively adjusting instruction and remediating student misunderstanding. Research emphasizes the importance of recognizing misconceptions to facilitate meaningful learning and long-term retention (Stefanich and Rokusek, 1992; Wilcox and Zielinski, 1997; Riccomini, 2005; Stein et al., 2005; Schnepper and McCoy, 2013). In the context of mathematics instruction, attention to the mathematical details in children's strategies provides valuable insights into their understanding (Easley and Zwoyer, 1975; Brown and Burton, 1978; Carpenter et al., 1999, 2003; Lester, 2007). They also allow the teacher to be intentional in remediating the student's mistakes. Prior education research discusses multiple good practices in remediating student mistakes. These range from offering visual aids (CAST, 2018), adopting a Socratic approach (Lepper and Woolverton, 2002), or eliciting student thinking through questions (Loewenberg Ball and Forzani, 2009). Effective instructional strategies coincide with strong teacher-student relationships, which significantly contribute to instructional effectiveness, stu-

dent motivation and student engagement (Wentzel, 1997; Pianta et al., 2003; Robinson, 2022; Wentzel, 2022).

2.2 Automated Feedback for Educators

Recent advances in NLP provide teachers feedback on their classroom discourse and have been shown to be beneficial, cost-effective feedback tools (Samei et al., 2014; Donnelly et al., 2017; Kelly et al., 2018; Jensen et al., 2020; Demszky and Liu, 2023; Wang and Demszky, 2023; Demszky et al., 2023). For example, Suresh et al. (2021) provides feedback to teachers on their teaching moves, such as how frequently the teacher asks students to reason aloud. Jacobs et al. (2022) provides evidence that K-12 math teachers receive this kind of feedback positively. The development of LLMs such as GPT-4 has re-kindled excitement around autotutors in providing equitable access to high-quality education (Graesser et al., 2004; Rus et al., 2013; Litman, 2016; Hobert and Meyer von Wolff, 2019; OpenAI, 2023; Khan Academy, 2023). However, these models are known to unreliably solve math problems (Frieder et al., 2023) and, more broadly, hallucinate information (Ji et al., 2023). Having a human tutor in-the-loop is important for catching these undesirable responses and preventing them from being disseminated to students. We explore the potential for a collaborative approach where a model and a human tutor could together provide students with effective guidance, thereby overcoming the limitations of either and ensuring a high-quality learning experience.

3 The ReMath Benchmark

This section discusses the ReMath benchmark which focuses on the step-by-step remediation process of expert teachers. The benchmark has three core tasks: **Task A**: Infer the type of student error, **Task B**: Determine a response strategy and intention of that strategy, and **Task C**: Generate a response that incorporates the information. Figure 1 illustrates ReMath. The following subsections describe the categories under each task; due to space constraints, we provide examples of each category in Appendix B. The framework for ReMath emerged from an extensive co-development with math teachers, spanning several months of collaboration. We developed it with the intention that the framework is comprehensive, intuitive, and aligned with the process that teachers actually follow. For more details on the development process, please refer to Appendix A.

3.1 Task A: Infer the Type of Student Error

Identifying the student's errors is prerequisite to successful remediation (Easley and Zwoyer, 1975; Bamberger et al., 2010). Task A involves teachers to infer the most likely cause of mistake from context. Prior research—particularly in math teacher education—has often focused on topic-specific categories of misconceptions, such as Bamberger et al. (2010). Our approach intends to support tutors who are not necessarily content experts, therefore we instead define "error" as a student's degree of understanding. As such, our error categories are topic-agnostic descriptions of a student's understanding, and complement the topic-agnostic strategies in Task B.

`guess`: The student does not seem to understand or guessed the answer. This error type is characterized by expressions of uncertainty or answers unrelated to the problem or target answer.

`misinterpret`: The student misinterpreted the question. This error type is characterized by answers that arise from a misunderstanding of the question. Students may mistakenly address a subtly different question, leading to an incorrect response. One example is reversing numbers, e.g., interpreting "2 divided by 6" as "6 divided by 2."

`careless`: The student made a careless mistake. This error type is characterized by answers that appear to utilize the correct mathematical operation but contain a small numerical mistake, resulting in an answer that is slightly off.

`right-idea`: Student has the right idea, but is not quite there. This error type is characterized by situations where the student demonstrates a general understanding of the underlying concept but falls short of reaching the correct solution.[1] For example, a student with `right-idea` error may recognize that multiplication is required to compute areas but may struggle with applying it to a specific problem.

`imprecise`: Student's answer is not precise enough or the tutor is being too picky about the

[1]This category is different from `careless` in that students with `right-idea` errors have difficulty in applying the concept correctly, whereas students with `careless` apply the concept correctly but make a minor numerical mistake.

form of the student's answer. This error type is characterized by answers that lacks the necessary level of precision or when the tutor places excessive emphasis on the form of the student's response.

`not-sure`: Not sure, but I'm going to try to diagnose the student. This option is used if the teacher is not sure why the student made the mistake from the context provided. We encourage the teachers to use this error type sparingly.

`N/A`: None of the above, I have a different description. This option is used when none of the other options reflect the error type. Similar to `not-sure`, we encourage teachers to use this error type sparingly.

3.2 Task B: Determine a Response Strategy and Intention of the Strategy

Student errors are usually persistent unless the teacher intervenes pedagogically (Radatz, 1980). This task involves selecting a course of action that guides the student towards improving their understanding. It also involves specifying the intention.

Strategies: Explain a concept, Ask a question, Provide a hint, Provide a strategy, Provide a worked example, Provide a minor correction, Provide a similar problem, Simplify the question, Affirm the correct answer, Encourage the student, Other.

Intentions: Motivate the student, Get the student to elaborate their answer, Correct the student's mistake, Hint at the student's mistake, Clarify a student's misunderstanding, Help the student understand the lesson topic or solution strategy, Diagnose the student's mistake, Support the student in their thinking or problem-solving, Explain the student's mistake (e.g., what is wrong in their answer or why is it incorrect), Signal to the student that they have solved or not solved the problem, Other.

3.3 Task C: Generate the Response

Once the student error has been identified and a response strategy has been determined, the next task is to generate a suitable response. We instruct teachers to respond in a useful and caring way. Experienced educators possess the instructional expertise to generate responses that are tailored to the individual student's needs (e.g., their error type) and age group. This is important as the students from this tutoring program are elementary school students, who require different pedagogical strategies than older students (Anghileri, 2006).

4 `ReMath` Formalism

This section presents the formalism for the `ReMath` benchmark. Given a conversation history c_h that contains evidence of a student's mistake, the ultimate goal is to generate a remediation response c_r that is useful and caring. The motivation behind `ReMath` is that experienced educators infer the student's error e and determine their response strategy and intention z before generating their final response. We model their response as c_r^*:

$$\underbrace{c_r^* \sim p(c_r|c_h, \underbrace{e}_{\text{Task A}}, \underbrace{z}_{\text{Task B}})}_{\text{Task C}}$$

In our dataset, a single annotation tuple is (c_h, e, z, c_r^*). Each tuple contains the conversation history c_h which includes the lesson topic and the last 5 conversation turns leading up to the student's turn where the mistake is made; in other words $c_h[-1]$ is the student's conversation turn where they make a mistake. Finally, the tuple contains the remediation annotations from the `ReMath` benchmark.

5 Dataset

Data source. Our data is sourced from a tutoring provider that offers end-to-end services for districts, including the tutoring platform, instructional materials, and tutors. The conversations used in `ReMath` were collected as part of their work in a Southern district in the United States that serves over 30k students. The students in these tutoring sessions are in the first to fifth grade, learning a variety of topics in math. The majority of schools are classified as Title I and three-quarters of students identify as Hispanic/Latinx. This district was focused on addressing existing achievement gaps among their students, as well as responding to the learning disruptions caused by the pandemic. To support the district's goals, the tutoring provider delivered 1:1 virtual tutoring to students at least twice a week, in-school over the course of the 2021-22 school year. The tutoring interactions are text-based, integrated on the providers' online platform. The platform has several features, including a whiteboard and pre-defined problems. The tutor communicates primarily through text message in a chat box, while the student uses either voice recording or the chat.

Preprocessing. The chat transcripts are de-identified by the tutoring provider. The student's

name is replaced with [STUDENT] and the tutor's name is replaced with [TUTOR]. ReMath uses excerpts from the original tutoring chat sessions, where the tutor responds to a mistake. Tutors on this platform use templated responses to flag mistakes, such as "That is incorrect" or "Good try." We leverage these templates to create a set of signalling expressions used by the tutor to identify excerpts. Specifically, we search for a three turn conversation pattern where (1) the tutor sends a message containing a question mark "?", (2) the student responds via text, then (3) the tutor uses a signalling expression. The set of signalling expressions were validated on a random sample of 100 conversations to ensure complete coverage. Appendix C includes the full set of signalling expressions we use.

Teacher annotation. We work closely with four math teachers from diverse demographics in terms of gender (3 female, 1 male) and race (Asian, Black/African American, White/Caucasian, Multiracial/Biracial). Three have more than 8 years of teaching experience, and the other has 6 years of teaching experience. They also have taught in a broad range of school settings including public schools, Title 1 schools, and charter schools. We work with two of the teachers in developing the framework and compensate them at $50/hour. We work with all four teachers in annotating the dataset, and compensate them at $40/hour.

Appendix A discusses the quality checks and onboarding process conducted prior to annotation. We randomly sampled 350 unique excerpts for annotation and assigned each to two teachers. Each teacher annotated whether the conversation has enough context about the problem and then added respective annotations for each of the three tasks.

Dataset statistics. The final dataset contains 700 items which we split into a train, validation, and test set with a 6:1:3 ratio. Each conversation excerpts contains 4 conversation turns. The train set contains 420, validation 70, and test 210 items.

6 Experiments

6.1 Models

We compare the teacher-written ReMath responses to four models. We fine-tune an instruction-tuned model Flan-T5 (large) (Chung et al., 2022). We also fine-tune a goal-directed dialog model GODEL (large) (Peng et al., 2022) because our data involves dialog-based interactions. Both models are fine-tuned on the training dataset using the teacher-generated responses, and not the original tutor responses. Appendix D contains details on the models and finetuning setup. We also ran GODEL and Flan-T5 zero-shot, however the generations were very poor upon manual inspection and have been omitted from the paper. We additionally compare against `gpt-3.5-turbo`[2] and `gpt-4` that have been optimized for chat. Appendix E contains the prompts used for `gpt-3.5-turbo` and `gpt-4` on the tasks. We use greedy decoding for all models.

6.2 Task Setup and Ablations

Task A. We prompt the models to predict the student's error type from prior context c_h: $\arg\max_e p(e|c_h)$.

Task B. We prompt the models to predict what strategy and intention to use from the context: $\arg\max_z p(z|c_h)$. Although there are many ways to predict z from context—for example, z could be predicted from e and c_h—our experiments focus on the setting where the strategy and intention are determined from c_h.

Task C. We prompt the models to generate a remediation response based on prior context, the error type and strategy : $c_r \sim p(c_r|c_h, e, z)$. We refer to the above as *complete-remediation generation*. In our evaluations, we provide the models with the teachers' error and strategy annotations. There are alternative comparisons such as prompting each model to generate their own error and strategy, then decoding their response from these predictions. We leave this extension for future work. We additionally run ablations to determine the importance of providing the error type and strategy. We run an *unconstrained generation* where the response is generated conditioned only on the context ($c_r \sim p(c_r|c_h)$), an *error-constrained generation* conditioned on the error type ($c_r \sim p(c_r|c_h, e)$) and a *strategy-constrained generation* conditioned on the strategy and intention annotation $c_r \sim p(c_r|c_h, z)$. Appendix E contains the prompts used for these settings.

6.3 Evaluation Setup

For both Task A and Task B, we want to measure the similarity between human and model annotations. To this end, we report the inter-rater reliability (IRR) as measured by Cohen's kappa and

[2]We use the `gpt-3.5-turbo-0301` model, and not the recently updated `gpt-3.5-turbo-0613` version.

the percentage of exact label matches between the human and the model. We are also interested in whether the models can identify a diverse set of errors, as the teachers did during the development of the ReMath benchmark; for this, we report entropy and the annotation percentage per category.

For Task C, we measure the extent to which the generated responses improve over the original tutors' responses. We recruit new teachers on Prolific (identified through Prolific's screening criteria) to perform pairwise comparisons between the original tutor response and a response generated by the teacher or one of the 16 models. A random set of 40 pairs per model is evaluated by 3 annotators each, who are blind to the source of the responses.

Raters evaluate the pairs along four dimensions. The first two items are *usefulness* and *care*, as these have been identified as key qualities of effective remediation in prior work (Roorda et al., 2011; Pianta, 2016; Robinson, 2022). The third item is *human-soundingness*; we discovered in a preliminary analysis of the data that low learning outcomes strongly correlated with whether the student was distracted by whether their tutor was a robot during their tutoring session. Given that the tutoring is chat-based, we include this as another dimension for measuring the effectiveness of responses. Each dimension is rated on a 5-point Likert scale with the following choices: Response A is much more <dimension> (caring/useful/human-sounding), Response A is somewhat more <dimension>, Response A and B are equally <dimension>, Response B is much more <dimension>, Response B is somewhat more <dimension>, where Response A represents the original tutor response and Response B represents the response by the teacher or the models. Finally, we ask the teachers to rate which responses they would *prefer* using, if they were the tutor. We use a similar 5-point Likert scale as the dimensions above indicating preference. We convert all Likert scale responses to integers between -2 and 2 for analysis. We additionally run automated metrics to compare the responses; however, we found these metrics not to be as insightful into the response quality and include them in Appendix G.

7 Results

Task A: Inference of error type. Table 1 summarizes the results. The teachers and the models all commonly annotate the student's error as guess. gpt-4 identifies the most diverse error categories and maintains the highest agreement with teachers out all the models. However, the relatively higher human IRR (0.38) and lower human-model IRR ($0.04 - 0.23$) indicate that there are still settings where the teachers and models disagreement on. Additionally, we find that the other models generally exhibit low diversity in selecting error type (rf. "entropy" column). The fine-tuned models potentially exhibit this behavior because the distribution of teacher-annotated errors is already skewed. We hypothesize that gpt-3.5-turbo exhibits low diversity because some categories like right-idea require knowledge of younger students' error traces, which may not be common in the model's training data.

We found that the human IRR being only fair (0.38) can be mostly attributed to some conversations not providing enough context about the conversation or context on the problem; for example, the tutor might flag a student's multiple choice answer as incorrect, however the conversation does not provide what options are available as the problem is being presented on their shared whiteboard screen. Nonetheless, we assume that even if those problems were available in the chat, the relative ordering of the IRR values would remain the same.

Task B: Identifying the strategy and intention. Two key observations can be made from Table 2. One, humans have lower agreement with each other on the strategies on this task compared to Task A. From our discussions with the teachers, the low agreement is due to them taking different strategies to remediate the same type of error. Employing different strategies is vital because it provides the student with multiple access points to understanding the content (Rose and Strangman, 2007; Glass et al., 2013; CAST, 2018).

This leads to our second observation which is that while teachers use a diversity of strategies, the models do not, as indicated by the models' lower entropy scores. Appendix G complement these observations with additional results on the strategy-intention pairings. These highlight how humans pick different strategies for the same intention (and vice versa), however models do not.

Task C: Generate response. Table 3 summarizes the results. Notably, models consistently outperform the original tutor response (indicated by positive values) on all dimensions, with the exception of strategy-constrained Flan-T5 being worse

	guess	misinterpret	careless	right-idea	imprecise	not-sure	N/A	entropy	match	IRR
human	**30%**	7%	11%	17%	8%	18%	9%	1.83	48%	0.38
Flan-T5 (f)	**87%**	0%	0%	0%	0%	12%	1%	0.44	33%	0.05
GODEL (f)	**50%**	0%	0%	15%	1%	30%	4%	1.18	31%	0.11
GPT-3.5-turbo	**34%**	33%	33%	0%	0%	0%	0%	1.10	20%	0.04
GPT-4	**37%**	21%	15%	2%	5%	20%	0%	1.53	37%	0.23

Table 1: Values show percentage of examples annotated with a given error type. "(f)" denotes the finetuned models. The most frequent error type for each model is **bolded**. The "match" column reports the % of items with an exact match with the human teacher label. IRR computes the Cohen kappa between the model and the human teachers (values are computed separately for each annotator then averaged across annotators).

	explain	ask	hint	strategy	example	correction	similar	simplify	affirm	encourage	N/A	entropy	match	IRR
human	18%	**30%**	8%	8%	3%	9%	7%	3%	6%	3%	5%	2.11	26%	0.12
Flan-T5 (f)	7%	22%	**71%**	0%	0%	0%	0%	0%	0%	0%	0%	0.75	14%	0.01
GODEL (f)	0%	43%	**57%**	0%	0%	0%	0%	0%	0%	0%	0%	0.68	21%	0.05
GPT-3.5	6%	15%	5%	2%	0%	**68%**	0%	2%	0%	1%	1%	1.12	14%	0.01
GPT-4	5%	**84%**	0%	0%	2%	3%	1%	0%	1%	2%	2%	0.73	31%	0.08

(a) Strategy

	correct E	hint E	diagnose E	explain E	clarify U	help U	elaborate A	support	motivate	signal	N/A	entropy	match	IRR
human	12%	14%	4%	1%	8%	**31%**	11%	6%	3%	5%	5%	2.08	13%	0.02
Flan-T5 (f)	0%	**71%**	17%	0%	0%	7%	5%	0%	0%	0%	0%	0.87	17%	0.04
GODEL (f)	0%	**57%**	0%	0%	0%	0%	43%	0%	0%	0%	0%	0.68	12%	0.0
GPT-3.5	**60%**	5%	0%	1%	8%	2%	13%	1%	3%	7%	0%	1.39	15%	0.04
GPT-4	3%	0%	**71%**	0%	17%	3%	1%	0%	2%	1%	2%	1.03	9%	0.03

(b) Intention

Table 2: Annotation results for the *strategy* (a) and *intention* (b). The results include the percentage of examples annotated with the given label, entropy of the label distribution, and agreement with human annotations ("match" and "IRR" columns, same as Table 1). "(f)" denotes the finetuned models. The most frequent label within a model is **bolded**. The Intention columns use a shortform for error (E), student's understanding (U), and answer (A).

on all dimensions and gpt-3.5-turbo being worse on care. The teacher-written responses are the most highly rated on all dimensions except, surprisingly, on human-soundingness.

The best model across all settings is gpt-4, and it benefits most from teacher-annotations. The ratings for gpt-4 from the *unconstrained* to *strategy-constrained* setting nearly double across all the dimensions except care. Its care rating only improves when error information is added. Figure 2 shows responses from gpt-4 that illustrate the diversity in remediation strategies. In the unconstrained setting, gpt-4 directly corrects the student, while the other models utilize different approaches to prompt the student to try again. The error-constrained gpt-4 provides a solution strategy tailored to the specific problem, while the strategy-constrained gpt-4 abstracts the details of applying the strategy. The complete-constrained gpt-4 model combines both approaches, offering a comprehensive but long remediation response. These results highlight the challenge for models in generating simultaneously useful and caring responses to student mistakes. Appendix G includes additional qualitative examples.

Unlike gpt-4, gpt-3.5-turbo's performance hurts with strategy information alone, however benefits from both having the error provided or both the error and strategy (rf. "overall"). We found that the the model's *strategy-constrained* generations tend to interpret the provided strategies in a corrective manner (e.g., immediately tells the correct answer) whereas the *error-constrained* generations propose alternative approaches (e.g., "try out this solution approach"). Table 5 illustrates this: the *strategy-constrained* response provides the worked out solution to the problem as the strategy annotation suggests, whereas the *error-constrained* response suggests the student to draw out a number line and count down.

The models fine-tuned on the expert teacher's data—Flan-T5 and GODEL—generally perform the best without constraints. These results suggest that Flan-T5 and GODEL are not good at generalizing

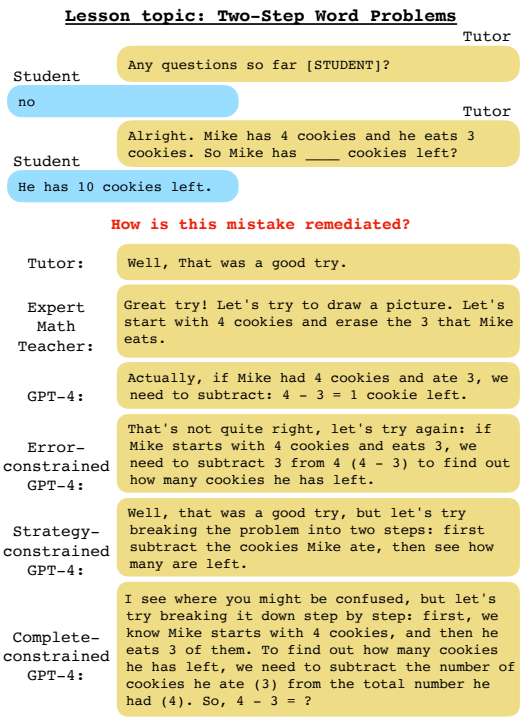


Figure 2: An example of how the tutor, a math teacher and gpt-4 models remediate this student's mistake. The error used here is guess, and the strategy and intention are to "Provide a solution strategy" and "Help student understand the lessons topic or solution strategy."

		Method	Prefer	Useful	Care	Not Robot	Overall
		Human	**1.26**	**1.19**	**0.86**	0.78	**1.02**
unconstrained	$p(c_r \mid c_h)$	Flan-T5	0.38	0.38	0.56	0.46	0.45
		GODEL	0.51	0.47	0.38	0.39	0.44
		GPT-3.5	0.46	0.45	−0.04	0.22	0.27
		GPT-4	0.54	0.54	0.50	0.47	0.51
error	$p(c_r \mid c_h, e)$	Flan-T5	0.17	0.17	0.17	0.10	0.15
		GODEL	0.23	0.24	0.26	0.40	0.28
		GPT-3.5	0.41	0.44	0.14	0.17	0.29
		GPT-4	0.88	0.64	0.79	0.83	0.79
strategy	$p(c_r \mid c_h, z)$	Flan-T5	−0.13	−0.15	−0.04	−0.03	−0.09
		GODEL	0.34	0.29	0.33	0.55	0.37
		GPT-3.5	0.27	0.29	−0.03	0	0.13
		GPT-4	0.97	1.08	0.5	**1.07**	0.91
complete	$p(c_r \mid c_h, e, z)$	Flan-T5	−0.02	0.11	0.11	0.16	0.09
		GODEL	0.38	0.23	0.45	0.88	0.48
		GPT-3.5	0.65	0.58	−0.04	0.59	0.45
		GPT-4	0.95	0.97	0.7	0.7	0.83

Table 3: Human evaluations on remediation responses written by educators (Human row) and models. The educator-written responses are grayed as a reference. The bolded values are the highest values within that column. The yellow cells are the highest values amongst all the models.

with constraints. This may be due to the imbalance of student error types and strategies in the training data, and the models' pre-training data. Human evaluators on Flan-T5 and GODEL frequently comments that the tutor responses were almost always factually incorrect. For example, when suggesting a worked out problem, *strategy-constrained* Flan-T5 generates a response like "Great try! Let's try another problem. Mike has 4 cookies and eats 3 cookies, so he has 4 - 3 = 2 cookies left." This contributes to the lower scores on the constrained generation setting.

8 Discussion & Conclusion

Our work focuses on remediation of student mistakes because mistakes are prime learning opportunities to correct misunderstandings that hinder learning progress. We present several contributions for understanding and performing remediation of student mistakes. First, we develop the ReMath framework, which sheds light on the thought process of experienced teachers for addressing student mistakes. This framework provides concrete tasks and evaluations for measuring the effectiveness of remediation responses. Second, we contribute a dataset that contains rich annotations from experienced math teachers, including the type of student error and the strategies the teachers would use to address the student's mistake. The dataset comes from a tutoring program working with a majority of Title I schools. We hope that it can serve as a valuable resource for providing equitable, high-quality learning experiences. Finally, we perform a thorough evaluation of responses from the experienced math teachers, instruct-tuned and dialog models on the ReMath tasks. We demonstrate that LLMs alone struggle to accurately infer student errors and generate useful, caring responses. However, when combined with the expert teacher annotations, the quality of LLM-generated responses significantly improves.

Our results indicate two promising avenues for scaling the remediation process. One approach is to directly prompt the model and ask human tutors to adapt the model's response such that they are appropriate for the student. Another approach is to prompt the tutor to identify the error and select a strategy from our predefined list, which can then be fed into the model and edited by tutor for further improvements. Our work shows promising results that indicate that either approach can lead to significant improvements over the tutor's response alone.

9 Limitations and Future Work

While our work provides a useful starting point for remediating student mistakes, there are limitations to our work. Addressing these limitations will be an important area for future research.

Access to questions. In some cases, the chat transcripts do not include the question the tutor and the student are working on together. This is because the questions are sometimes displayed on a shared whiteboard, and not posted in the chat. Even though our dataset includes annotations for when there's not enough context, future work could improve upon our analysis by always including information about the question. For example, this may improve the IRR scores on Task A.

Expanding to other subjects. Our dataset and benchmark currently focuses on mathematics. The taxonomy may not directly transfer to other subjects, although it may serve as a good starting point for remediating student mistakes in other domains.

Evaluation with students. Our human evaluations are currently limited to the teacher's perspective. However, ultimately, the effectiveness of the responses relies on how students receive and interpret them, and whether these interactions positively impact their learning outcomes. To address this limitation, future research should work towards evaluating this method with students. This is important as previous studies like Wentzel (2022) highlight the potential disparity between teachers and students in determining what responses are more caring or useful.

Ethics Statement

We recognize that our research on the integration of large language models (LLMs) in education ventures into a less explored territory of NLP with numerous ethical considerations. LLMs open up new possibilities for enhancing the quality of human education, however there are several ethical considerations we actively took into consideration while performing this work. We hope that these serve as guidelines for responsible practices, and hope that future work does the same.

First is the privacy of both students and tutors. We obtained approval from the tutoring program for repurposing the data into the `ReMath` dataset. We handled all data with strict confidentiality, adhering to best practices in data anonymization and storage security.

Furthermore, we are committed to promoting equity and inclusivity in education. The compensation provided to the experienced math teachers involved in our benchmarking process was set at a significantly higher rate, reflecting our recognition of their invaluable contributions and domain expertise. By compensating teachers fairly, we aim to foster a culture of respect, collaboration, and mutual support within the NLP and education community.

Finally, we are committed to the responsible use of our research findings. We encourage the adoption of our benchmark and methodologies by the research community, with the understanding that the ultimate goal is to improve educational outcomes for all students and provide support to educators. We actively promote transparency, openness, and collaboration to drive further advancements in the field of natural language processing (NLP) for education.

References

National Student Support Accelerator. 2022. Using the American Rescue Plan Act Funding For High-Impact Tutoring. https://studentsupportaccelerator.org/briefs/using-american-rescue-plan. [Online; accessed 4-June-2023].

Julia Anghileri. 2006. Scaffolding practices that enhance mathematics learning. *Journal of Mathematics Teacher Education*, 9:33–52.

Honi Joyce Bamberger, Christine Oberdorf, and Karren Schultz-Ferrell. 2010. *Math misconceptions: PreK-grade 5: From misunderstanding to deep understanding*. Heinemann.

Jo Boaler. 2013. Ability and mathematics: The mindset revolution that is reshaping education. Forum.

John Seely Brown and Richard R Burton. 1978. Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive science*, 2(2):155–192.

Thomas P Carpenter, Elizabeth Fennema, M Loef Franke, Linda Levi, and Susan B Empson. 1999. Children's mathematics. *Cognitively Guided*, 8.

Thomas P Carpenter, Megan Loef Franke, and Linda Levi. 2003. *Thinking mathematically*. Portsmouth, NH: Heinemann.

CAST. 2018. Universal Design for Learning Guidelines version 2.2. Retrieved from http://udlguidelines.cast.org. https://udlguidelines.cast.org/. [Online; accessed 4-June-2023].

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Dorottya Demszky and Jing Liu. 2023. M-powering teachers: Natural language processing powered feedback improves 1:1 instruction and student outcomes. *L@S '23: Proceedings of the Tenth ACM Conference on Learning @ Scale*.

Dorottya Demszky, Jing Liu, Heather Hill, Dan Jurafsky, and Chris Piech. 2023. 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*.

P. J. Donnelly, N. Blanchard, A. M. Olney, S. Kelly, M. Nystrand, and S. K. D'Mello. 2017. Words matter: Automatic detection of teacher questions in live classroom discourse using linguistics, acoustics and context. 218–227. Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK '17.

J Al Easley and Russell E Zwoyer. 1975. Teaching by listening-toward a new day in math classes. *Contemporary Education*, 47(1):19.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. 2023. Mathematical capabilities of chatgpt. *arXiv preprint arXiv:2301.13867*.

Roland G Fryer Jr and Meghan Howard-Noveck. 2020. High-dosage tutoring and reading achievement: evidence from new york city. *Journal of Labor Economics*, 38(2):421–452.

Don Glass, Anne Meyer, and David Rose. 2013. Universal design for learning and the arts. *Harvard Educational Review*, 83(1):98–119.

Arthur C Graesser, Shulan Lu, George Tanner Jackson, Heather Hite Mitchell, Mathew Ventura, Andrew Olney, and Max M Louwerse. 2004. Autotutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, 36:180–192.

Sebastian Hobert and Raphael Meyer von Wolff. 2019. Say hello to your new automated tutor–a structured literature review on pedagogical conversational agents.

Jennifer Jacobs, Karla Scornavacco, Charis Harty, Abhijit Suresh, Vivian Lai, and Tamara Sumner. 2022. Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change. *Teaching and Teacher Education*, 112:103631.

E. Jensen, M. Dale, P. J. Donnelly, C. Stone, S. Kelly, A. Godley, and S. K. D'Mello. 2020. Toward automated feedback on teacher discourse to enhance teacher learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

S. Kelly, A. M. Olney, P. Donnelly, M. Nystrand, and S. K. D'Mello. 2018. Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, 47:7.

Sean Kelly, Robert Bringe, Esteban Aucejo, and Jane Cooley Fruehwirth. 2020. Using global observation protocols to inform research on teaching effectiveness and school improvement: Strengths and emerging limitations. *Education Policy Analysis Archives*, 28:62–62.

Khan Academy. 2023. Harnessing GPT-4 so that all students benefit. A nonprofit approach for equal access. `https://blog.khanacademy.org/harnessing-ai-so-that-all-students-benefit-a-nonprofit-approach-for-equal-access`. [Online; accessed 4-June-2024].

M. A. Kraft, D. Blazar, and D. Hogan. 2018. The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, 88(4):547–588.

Mark R Lepper and Maria Woolverton. 2002. The wisdom of practice: Lessons learned from the study of highly effective tutors. In *Improving academic achievement*, pages 135–158. Elsevier.

Frank K Lester. 2007. *Second handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics*. IAP.

MiniChain Library. 2023. MiniChain Library. `https://github.com/srush/minichain#typed-prompts`. [Online; accessed 4-June-2024].

Diane Litman. 2016. Natural language processing for enhancing teaching and learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Deborah Loewenberg Ball and Francesca M Forzani. 2009. The work of teaching and the challenge for teacher education. *Journal of teacher education*, 60(5):497–511.

Ian McKenzie. 2023. Inverse Scaling Prize: First Round Winners. https://irmckenzie.co.uk/round1#:~:text=model%20should%20answer.-,Using%20newlines,-We%20saw%20many. [Online; accessed 4-June-2024].

Andre Nickow, Philip Oreopoulos, and Vincent Quan. 2020. The impressive effects of tutoring on prek-12 learning: A systematic review and meta-analysis of the experimental evidence.

U.S. Department of Education. 2021. Strategies for Using American Rescue Plan Funding to Address the Impact of Lost Instructional Time. https://www2.ed.gov/documents/coronavirus/lost-instructional-time.pdf. [Online; accessed 4-June-2023].

OpenAI. 2023. Gpt-4 technical report.

Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. 2022. Godel: Large-scale pre-training for goal-directed dialog. arXiv.

Robert C Pianta. 2016. Teacher–student interactions: Measurement, impacts, improvement, and policy. *Policy insights from the behavioral and brain sciences*, 3(1):98–105.

Robert C Pianta, Bridget Hamre, and Megan Stuhlman. 2003. Relationships between teachers and children.

Hendrik Radatz. 1980. Students' errors in the mathematical learning process: a survey. *For the learning of Mathematics*, 1(1):16–20.

Paul J Riccomini. 2005. Identification and remediation of systematic error patterns in subtraction. *Learning Disability Quarterly*, 28(3):233–242.

Carly D Robinson. 2022. A framework for motivating teacher-student relationships. *Educational Psychology Review*, 34(4):2061–2094.

Carly D Robinson and Susanna Loeb. 2021. High-impact tutoring: State of the research and priorities for future learning. *National Student Support Accelerator*, 21(284):1–53.

Debora L Roorda, Helma MY Koomen, Jantine L Spilt, and Frans J Oort. 2011. The influence of affective teacher–student relationships on students' school engagement and achievement: A meta-analytic approach. *Review of educational research*, 81(4):493–529.

David H Rose and Nicole Strangman. 2007. Universal design for learning: Meeting the challenge of individual learning differences through a neurocognitive perspective. *Universal access in the information society*, 5:381–391.

Vasile Rus, Sidney D'Mello, Xiangen Hu, and Arthur Graesser. 2013. Recent advances in conversational intelligent tutoring systems. *AI magazine*, 34(3):42–54.

B. Samei, A. M. Olney, S. Kelly, M. Nystrand, S. D'Mello, N. Blanchard, X. Sun, M. Glaus, and A. Graesser. 2014. Domain independent assessment of dialogic properties of classroom discourse.

Lauren C Schnepper and Leah P McCoy. 2013. Analysis of misconceptions in high school mathematics. *Networks: An Online Journal for Teacher Research*, 15(1):625–625.

Meghan Shaughnessy, Rosalie DeFino, Erin Pfaff, and Merrie Blunk. 2021. I think i made a mistake: How do prospective teachers elicit the thinking of a student who has made a mistake? *Journal of Mathematics Teacher Education*, 24:335–359.

Greg P Stefanich and Teri Rokusek. 1992. An analysis of computational errors in the use of division algorithms by fourth-grade students. *School Science and Mathematics*, 92(4):201.

Marcy Stein, Diane Kinder, Jerry Silbert, and Douglas W Carnine. 2005. *Designing effective mathematics instruction: A direct instruction approach*. Pearson.

A. Suresh, J. Jacobs, V. Lai, C. Tan, W. Ward, J. H. Martin, and T. Sumner. 2021. Using transformers to provide teachers with personalized feedback on their classroom discourse: The talkmoves application. arxiv. Preprint.

Rose Wang and Dorottya Demszky. 2023. Is chatgpt a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. In *18th Workshop on Innovative Use of NLP for Building Educational Applications*.

Rose Wang, Pawan Wirawarn, Noah Goodman, and Dorottya Demszky. 2023. Sight: A large annotated dataset on student insights gathered from higher education transcripts. In *Proceedings of Innovative Use of NLP for Building Educational Applications*.

Kathryn R Wentzel. 1997. Student motivation in middle school: The role of perceived pedagogical caring. *Journal of educational psychology*, 89(3):411.

Kathryn R Wentzel. 2022. Does anybody care? conceptualization and measurement within the contexts of teacher-student and peer relationships. *Educational Psychology Review*, pages 1–36.

Sandra K Wilcox and Ronald S Zielinski. 1997. Implementing the assessment standards for school mathematics: Using the assessment of students' learning to reshape teaching. *The Mathematics Teacher*, 90(3):223–229.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*.

A Developing the ReMath Benchmark

This section details how we developed the ReMath Benchmark in collaboration with the math teachers. The design objective of the benchmark is to capture the teachers' thought process when addressing student mistakes. We developed the taxonomy closely with two of the four teachers. We compensated them at $50/hour. We met with them on a weekly to biweekly basis.

During the preliminary stages of this work, we provided the teachers examples of the conversations and asked them to directly revise the tutor's responses. For the first few weeks, we met on a weekly basis where a co-author presented the teachers about 20 conversation examples and the teachers worked on the examples asynchronously. During the meetings, the teachers and co-author discussed the teachers' approaches to the setting. After four meetings, themes started to emerge in the types of approaches the teachers used. For instance, the teachers often made hypotheses about the student's thought process, which gave rise to the error category. This illustrated that educators possess a mental model of what the student is doing and employ various probing techniques to confirm or refute their hypotheses. The diverse ways in which the teachers probed and engaged with the students led to the identification of different strategies. We further categorized these strategies based on their intentions, reflecting the potential consequences they might have on the student's learning process.

We then created a taxonomy of these approaches, and edited the taxonomy through more iterations of task attempts and discussion. These edits included expanding the set of categories, removing irrelevant categories, separating categories into different groups (e.g., the separation of student error from the teacher's strategies) and re-structuring the order of the tasks. The taxonomy was finalized once both teachers and the co-authors were satisfied with how naturally the benchmark could be used and with the benchmark's coverage.

B Category examples in ReMath

This section provides examples for each of the categories in the ReMath benchmark. It is split by *error type*, *response strategy*, and *intention*.

B.1 Student Error Types

guess: The student does not seem to understand or guessed the answer. This error type is characterized by expressions of uncertainty or answers that do not seem related to the problem, the options or the target answer. An example of this is the following conversation snippet on the topic of "Addition and subtraction within 100":
tutor: We need to subtract 6 from 15.
student: oh okay...
tutor: What is the value of 15 - 6?
student: it is 11?
This example could be labeled as the student guessing because they express uncertainty in their answer ("it is 11?")

misinterpret: The student misinterpreted the question. This error type is characterized by answers that arise from a misunderstanding of the question being asked. Students may mistakenly address a subtly different question, leading to an incorrect response. For example, a common manifestation of this error is the reversal of number orderings, such as interpreting "2 divided by 6" as "6 divided by 2." An example of this is the following conversation snippet on the topic of "Converting Units of Measure":
student: sorry for the j that I tipe.
tutor: Not an issue, [STUDENT].
tutor: How many times 1000 will goes into 7000?
student: it cant
This example could be labeled as the student misinterpreting because the student might have read the question as the reverse question (e.g., "How many times can 7000 go into 1000?") because they say that the number cannot go into the other number.

careless: The student made a careless mistake. This error type is characterized by answers that appear to utilize the correct mathematical operation but contain a small numerical mistake, resulting in an answer that is slightly off. It reflects a lack of careful attention to detail or a minor computational error in an otherwise sound solution approach. An example of this is the following conversation snippet on the topic of "Volume of Rectangular Prisms":
tutor: Again, we have to multiply the value of 6 with 20.
student: so it is 110
tutor: So, what is the value of 20 times 6?
student: 110
This example could be labelled as the student mak-

ing a careless mistake. The student seems capable of multiplying (their answer is larger than 100) and does not mistake the operation (e.g., they multiply, and do not add the numbers). They make a minor mistake in the calculation (110 instead of 120), which suggests that they made a careless mistake.

`right-idea`: Student has the right idea, but is not quite there. This error type is characterized by situations where the student demonstrates a general understanding of the underlying concept or approach but falls short of executing or reaching the correct solution. For example, a student may recognize that multiplication is required to compute areas but may struggle with applying it to a specific problem. An example of this is the following conversation snippet on the topic of "Area":
tutor: Please check the question once.
tutor: The factors are 24 and 86.
tutor: What is the formula for finding the area of a rectangle?
student: multiplying
tutor: So, what is the value of 20 times 6?
student: 110
This example could be labelled as the student having the right idea, but isn't quite there. The student seems to understand what operation is need for calculating the area, but their language is not precise (e.g., they don't mention 'width' or 'length'). This suggests that they might not have a clear understanding of how to apply the concept.

`imprecise`: Student's answer is not precise enough or the tutor is being too picky about the form of the student's answer. This error type is characterized by student answers that lacks the necessary level of precision or when the tutor places excessive emphasis on the specific form of the student's response. An example of this is the following conversation snippet on the topic of "Concept of Area":
student: yes
tutor: Okay!
tutor: What should he measure?
student: the dimensional area
In this example, the tutor flags the student's answer as incorrect, and says that the correct answer is "area." This example could be labelled by this error because the student either is imprecise with their language and/or the tutor is being too strict about the use of term.

`not-sure`: Not sure, but I'm going to try to diagnose the student. This option is used if the teacher is not sure why the student made the mistake from the context provided. We encourage the teachers to use the provided lesson topic and their teaching experience with students to determine what the mistake is, and use this error type sparingly.

`N/A`: None of the above, I have a different description. This option is used of none of the other options reflect the error type. Similar to `not-sure`, we encourage teachers to use this error type sparingly.

B.2 Response Strategies and Intentions

Below are examples of response strategies and intentions that the teachers selected. We provide the lesson topic to each example. The original tutor's messages are marked with *tutor*, and the students' with *student*. Note that in the annotation setup, we allow the teachers to simulate the student's response in order for the teachers to fully complete their strategy. Therefore, the examples here will include the teacher's simulated response for the student. The teacher's response is marked with *teacher*, and the simulated student messages are marked with *student (simulated)*. The teachers selected strategies that they used for their own response—the strategies are not descriptions of what the tutor did.

As a reminder, the categories of response strategies are:

- Explain a concept
- Ask a question
- Provide a hint
- Provide a strategy
- Provide a worked example
- Provide a worked example
- Provide a minor correction
- Provide a similar problem
- Simplify the question
- Affirm the correct answer
- Encourage the student
- Other

The categories of intentions are:

- Motivate the student
- Get the student to elaborate their answer
- Correct the student's mistake
- Hint at the student's mistake
- Clarify a student's misunderstanding
- Help the student understand the lesson topic or solution strategy
- Diagnose the student's mistake
- Support the student in their thinking or problem-solving
- Explain the student's mistake (e.g., what is wrong in their answer or why is it incorrect)
- Signal to the student that they have solved or not solved the problem
- Other (use sparingly).

Here are some example of strategies used, and their paired intention.

Explain a concept (Intention: Correct the student's mistake) *Lesson*: Order of Operations
tutor: You earn one point for your good beginning.
tutor: That was a good try.
tutor: What is the value of 3^3?
student: 9
teacher: When we see an exponent, we multiply the number by itself however much the exponent is. Is this example, that is 3 times 3 times 3. What is that value?
student (simulated): 27
teacher: Correct. 3 raised to the third power is 3 times 3 times 3, which is 27. Let's try another one.",

Ask a question (Intention: Correct the student's mistake) *Lesson*: Identify Unit Rate
student: ok
tutor: We can find the unit rate by selecting one point on the line and writing down the coordinates.
tutor: What is the coordinate of one point on the line?
student: 300
teacher: Great try! Remember that for coordinates we need an x and a y value. Which one is 300?
student (simulated): x
teacher: Great job - now what about y?

Provide a hint (Intention: Hint at the student's mistake) *Lesson*: Multiplying Multi-Digit by One-Digit Numbers
tutor: Give it your best shot.
tutor: Good going!
tutor: Is that your final answer?
student: yes
teacher: Great try! Remember when multiplying that we need to carry the tens place
student (simulated): o yah. So I carry the 2?
teacher: Exactly!

Provide a strategy (Intention: Support the student in their thinking or problem-solving) *Lesson*: Metric Conversions and Word Problems
student: "yes i understand
tutor: Let's move on to the next question.
tutor: How many milliliters are in one liter?
student: is that a 20
teacher: Great try! Try using the prefixes to help you. Milli means thousand.
student (simulated): oh then thousand?
teacher: "Yes! There are 1000 milliliters in a liter.

Provide a worked example (Intention: Help the student understand the lesson topic or solution strategy) *Lesson*: Understanding Multiplication
tutor: Go ahead and give your best.
student: c
tutor: Is that your final answer?
student: yes
teacher: That's a great try! When we are multiplying, try to think of x groups of y and how many that would leave you. For example, 3 times 5 would be 3 groups of 5 or 5 groups of 3, which is 15.
teacher: So, what would 2 times 4 be? *student (simulated)*: 8?
teacher: Great job!

Provide a minor correction (Intention: Support the student in their thinking or problem-solving) *Lesson*: Decomposing Figures
tutor: What is the area of the bottom rectangle?
tutor: That's correct
tutor: What is the area of the top rectangle?
student: es 50
teacher: Ah, not quite. What is the length and the width of the top rectangle?
student (simulated): 2 and 5

teacher: Yes. Do you remember how to get the area of a rectangle?
student (simulated): we multiply
teacher: Yes, please multiply 2 x 5
student (simulated): 10

Provide a similar problem (Intention: Help the student understand the lesson topic or solution strategy) *Lesson*: Dividing Using Equations
tutor: Try to answer my question.
tutor: Okay, let me explain it to you.
tutor: What is the value of 5 times 2?
student: 5
teacher: Great try! 5 times 1 is 5. What about 5 times 2?
student (simulated): twice. so 10
teacher: Correct!

Simplify the question (Intention: Help the student understand the lesson topic or solution strategy)

Affirm the correct answer (Intention: Signal to the student that they have solved or not solved the problem) *Lesson*: Shapes and Area
tutor: I was facing some technical glitch.
student: ya
tutor: What is area?
student: spase in side sompthing
teacher: Yes well done! Area is the 2D space inside something
student (simulated): ok
teacher: Surface area is the area on the outside of a 3D object

Encourage the student (Intention: Motivate the student) *Lesson*: Rounding
tutor: Here, the value of 6 in the given number is more than 5.
tutor: So, we need to round the value 7.
tutor: Do you have any questions about that part?
student: millions
teacher: Ok, ask away!
student (simulated): why do we round up?
teacher: Becuase the 6 is greater than 5 (5 is the cutoff)

C `ReMath` Dataset Collection

This section discusses how the initial dataset is processed and how the dataset is annotated.

C.1 Data Processing

Signalling Expressions for Student Mistakes The following is the list of the signalling expressions used by the tutor which we use to mark conversation segments where the student has made a mistake. To identify these segments, we first lowercase all the conversation utterances, and check whether the following expressions exactly occur in the conversation.

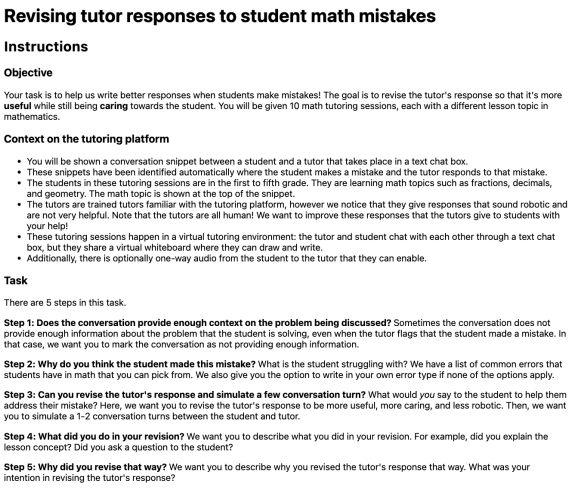
- "incorrect"
- "not quite"
- "bit off"
- "good try"
- "great try"
- "effort"
- "recheck"

C.2 Annotation Quality Check

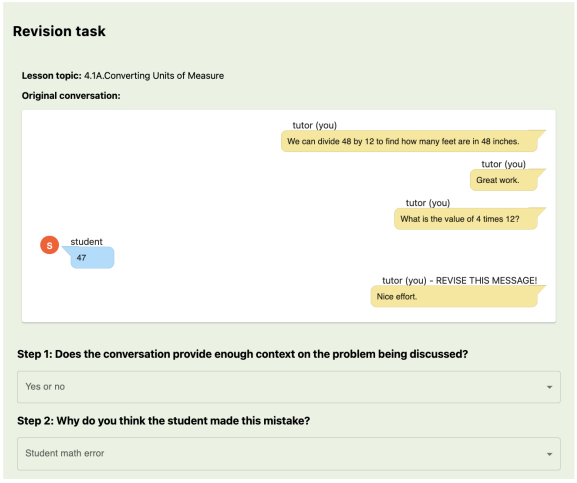
We perform quality checks before the teachers started annotation. First, they are onboarded by an author of this work through two meetings, each meeting ranging between 30-60 minutes. After the meeting, the teachers complete a sample of 20 problems similar to the ones in the final task. The teachers and author then meet again to walk through their answers and check their understanding of each of the taxonomy's category options. The 20 sample problems are not used for the dataset and are only for onboarding purposes. After training, each item took about 2 to 10 minutes for the teachers to complete.

C.3 Annotation Setup

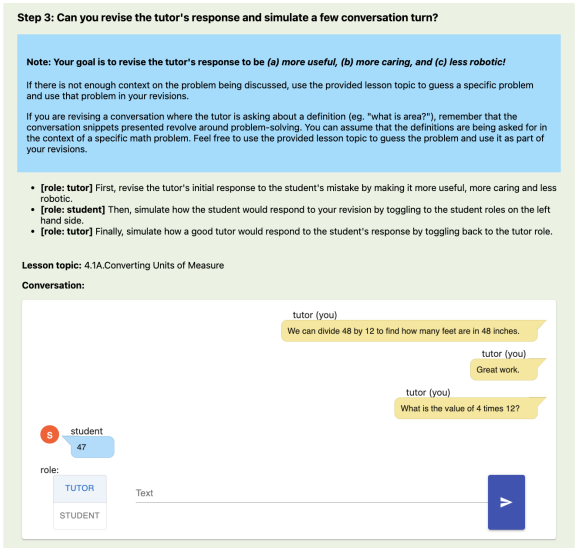
Figure 3 shows the interface used by the teachers for annotating the data in our `ReMath` dataset. Note that the annotation interface allows teachers to simulate the student's response. We have this feature because the teachers found that only responding on a single turn was not sufficient for them to complete their strategy of choice.



(a) Instructions



(b) Step 1 & 2



(c) Step 3

(d) Step 4 & 5

Figure 3: Annotation interface for collecting ReMath data.

D Models

Flan-T5: We fine-tune the Flan-T5-large checkpoint with 780M parameters based on the Huggingface repository version. The model is trained with a batch size of 8 sequences, a learning rate of 5e-5, and a weight decay of 0.01 for 5 training epochs. For Task A and Task B, we use dynamic padding for the batches; for Task C, we use a maximum sequence length of 512.

GODEL: We fine-tuned the GODEL large (770M parameters) based on the HuggingFace repository version, with 24 layers of encoder and decoder and 1024 dimensional embeddings. The hyperparameters and sequence lengths used are the same as those used in fine-tuning Flan-T5.

Each best model is selected based on the perplexity score on the validation set. We use beam search for inference using a beam size of 1.

E Prompts and Input Formatting

This section contains information on how the prompts and inputs are formatted for Flan-T5 and GODEL. It also includes information on the prompts used for both gpt-3.5-turbo and gpt4 on the tasks in the ReMath benchmark. Unless otherwise noted, our prompt practices follow a mix of works from NLP, education and social sciences (McKenzie, 2023; Library, 2023; Ziems et al., 2023; Wang et al., 2023). Specifically, for Task A and Task B, which involve categorical answers, we enumerate the options as multiple-choice options and each option is separated by a newline. We add context for the task at the start of the prompt, and the con-

straints of outputting a JSON-formatted text for the task at the end of the prompt. For Task A and Task B, we found it important to repeat the constraints at the end and specify the JSON format to ensure that the model pick one of the options provided. For Task C, we found it important to add a length constraint to force the model to stick to the short message styles of the tutor and student; otherwise, the model responses would generally be extremely long (up to $5-10\times$ longer than the original tutor responses). Adding the length constraint also prevented the model from simulating the rest of the tutoring session.

E.1 Task A: Infer student's error

The prompt used for `gpt-3.5-turbo` and `gpt-4` on Task A is shown in Figure 5.

To fine-tune `Flan-T5` and `GODEL` for Task A, we cast the classification task into a "text-to-text" format with the student error type label as the corresponding target. The detailed preprocessing formats for the input sequence for fine-tuning `Flan-T5` and `GODEL` can be seen in Figure 6 and Figure 7, respectively.

E.2 Task B: Determine the strategy and intention

The prompt used for `gpt-3.5-turbo` and `gpt-4` on Task B is shown in Figure 8. For fine-tuning `Flan-T5` and `GODEL`, similarly, we cast the classification task of categorizing strategies and intentions into "text-to-text" formats. The detailed preprocessing formats for the input sequence for Task B can be seen in Figure 9 and Figure 10, respectively.

E.3 Task C: Generate the remediation response

The prompt used for `gpt-3.5-turbo` and `gpt-4` on Task C is shown in Figure 11. The input sequence formats used for Flan-T5 and GODEL on Task C are shown in Figure 12 and Figure 13, respectively.

F Task C Human Evaluations

We describe the human evaluation setup for Task C, whose results are reported in Section 7.

The human evaluations were run on Prolific. Our prescreening criteria were that the participants have to be located in the USA, have to be a teacher, their fluent languages have to include English, and their approval rating has to be at least 96%. We conduct the human evaluations on 40 items from each model with 3 raters; 10 of these items were held to be the same and the other 30 were randomly sampled. The 10 items are used to calculated the IRR reported in the main tables. Each item consisted of a pair of remediation responses, Response A and Response B. One of the responses is the original tutor's response to the student's mistake, and the other response is the newly generated remediation response (ie. the expert-written response in the `Human` row, and the model-generated response in the other rows). The ordering of the responses is always randomized. Each item is scored on a Likert scale from -2 to 2 on four dimensions: *usefulness*, *care*, *human-soundingness*, and *preference*. We also provided a definition for each dimension.

Figure 14 shows an example of the evaluation interface. Specifically, the phrasing for each dimension was:

Which response is more useful?
Definition: Useful responses are responses that are productive at advancing the student's understanding and helping them learn from their errors. These are responses that lead to the student getting similar questions right in the future, and not just figuring out the answer to this specific problem.

- Response A is much more useful.
- Response A is somewhat more useful.
- Responses A and B are equally useful.
- Response B is somewhat more useful.
- Response B is much more useful.

Which response is more caring?
Definition: Caring responses are responses that express kindness or concern for the student. They foster a collaborative and supportive relationship between the tutor and the student.

- Response A is much more caring.
- Response A is somewhat more caring.
- Responses A and B are equally caring.
- Response B is somewhat more caring.
- Response B is much more caring.

Which response is more human-sounding?
Which of the responses sounds more human, and less like a machine or artificial intelligence entity typed it?

- Response A is much more human-sounding.
- Response A is somewhat more human-sounding.
- Responses A and B are equally human-sounding.
- Response B is somewhat more human-sounding.
- Response B is much more human-sounding.

Which response would you rather choose to respond with if you were the tutor?

- I strongly prefer to pick Response A.
- I prefer to pick Response A.
- I equally prefer either Response A or B.
- I prefer to pick Response B.
- I strongly prefer to pick Response B.

G Additional Results

G.1 Task A

Figure 4 shows the distribution over error categories selected by the teachers and the models.

G.2 Task B

Figure 4 shows the distribution over strategies and intentions categories selected by the teachers and the models.

Figure 15 shows the pairing of strategies and intentions selected by the models, and the teachers.

G.3 Task C

The following tables show examples of the model outputs.

- Table 4 shows examples for `gpt-4`.
- Table 5 shows examples for `gpt-3.5-turbo`.
- Table 6 shows examples for `Flan-T5`.
- Table 7 shows examples for `GODEL`.

Table 8 reports the automated metrics on the model and human generated responses.

Figure 16 illustrates an example of two teachers taking different strategies to remediate the same type of error. Employing different strategies is vital because it provides the student with multiple access points to understanding the content (Rose and Strangman, 2007; Glass et al., 2013; CAST, 2018).

c_h
Lesson topic: Rounding Whole Numbers ... *tutor*: Here you go, give it a shot. *student*: ok *tutor*: Do you need any hints to solve this one? *student*: 142

(a)

Method	**Response**
`Original`	You made a good try.
`Human`	That is a good try. Let's review place value. In the number, 142, what number is in the tens place?
`gpt-4`	That's a good attempt, but remember when rounding whole numbers, we need to look at the place value we're rounding to. Can you try again and round 142 to the nearest ten?
`gpt-4` (error)	I see that you got 142, but let's take a closer look at the rounding rules to make sure we're applying them correctly.
`gpt-4` (strategy)	That's close, but remember when rounding to the nearest ten, we look at the digit in the ones place; if it's 5 or more, we round up, and if it's 4 or less, we round down. Let's try another problem: round 137 to the nearest ten.
`gpt-4` (complete)	I see that you got 142, but let's take a closer look at the problem together and try another example: Round 138 to the nearest ten.

(b)

Table 4: (a) show the original context, and (b) shows the responses from the original tutor, teacher and `gpt-4` methods.

c_h
Lesson topic: Addition and Subtraction Within 100 ... *tutor*: We need to subtract 6 from 15. *student*: oh okay.. *tutor*: What is the value of 15 - 6? *student*: it is 11?

,

(a)

Method	**Response**
`Original`	That was a good effort.
`Human`	You're close! Let's do 24 - 7 on the whiteboard and then you can think about your answer again.
`gpt-3.5-turbo`	Great effort, but double check your answer - it should be 9.
`gpt-3.5-turbo` (error)	Great start, but double check your answer by counting backwards from 15 to 6.
`gpt-3.5-turbo` (strategy)	Great effort, but the correct answer is actually 9.
`gpt-3.5-turbo` (complete)	Great job! Let's try using a number line to help us subtract 6 from 15. First, draw a number line from 0 to 15. Then, start at 15 and count back 6 spaces. What number do you land on? That's the answer!

(b)

Table 5: (a) show the original context, and (b) shows the responses from the original tutor, teacher and `gpt-3.5-turbo` methods. In this example, the error identified is that the student has the `right-idea`. The strategy is to provide a worked example, and the intention is to get the student to understand the lesson topic or solution strategy.

c_h
Lesson topic: Two-Step Word Problems ... *tutor*: Any questions so far [STUDENT]? *student*: no *tutor*: Alright. Mike has 4 cookies and he eats 3 cookies. So Mike has _____ cookie left? *student*: He has 10 cookies left.

(a)

Method	**Response**
`Original`	Well, That was a good try. Plus one point for your responces.
`Human`	Great try! Let's try to draw a picture. Let's start with 4 cookies and erase the 3 that Mike eats.
`Flan-T5`	Great try! Let's try another problem. What number of cookies does Mike have left after he eats 3?
`Flan-T5` (error)	Great try! Let's try another problem. What number of cookies did Mike have before he ate 3?
`Flan-T5` (strategy)	Great try! Let's try another problem. Mike has 4 cookies and eats 3 cookies, so he has 4 - 3 = 2 cookies left.
`Flan-T5` (complete)	Great try! When you eat a cookie, you have to subtract the number of cookies from your original number. For example, if Mike has 4 cookies and eats 3, how many cookies are left?

(b)

Table 6: (a) show the original context, and (b) shows the responses from the original tutor, teacher and `Flan-T5` methods. In this example, the error identified is `guess`. The strategy is to provide a solution strategy, and the intention is to get the student to understand the lesson topic or solution strategy.

c_h
Lesson topic: Prime and Composite Numbers ... *student*: okay *student*: 34 *tutor*: 39 is a prime number or composite number? *student*: prime *student*: He has 10 cookies left.

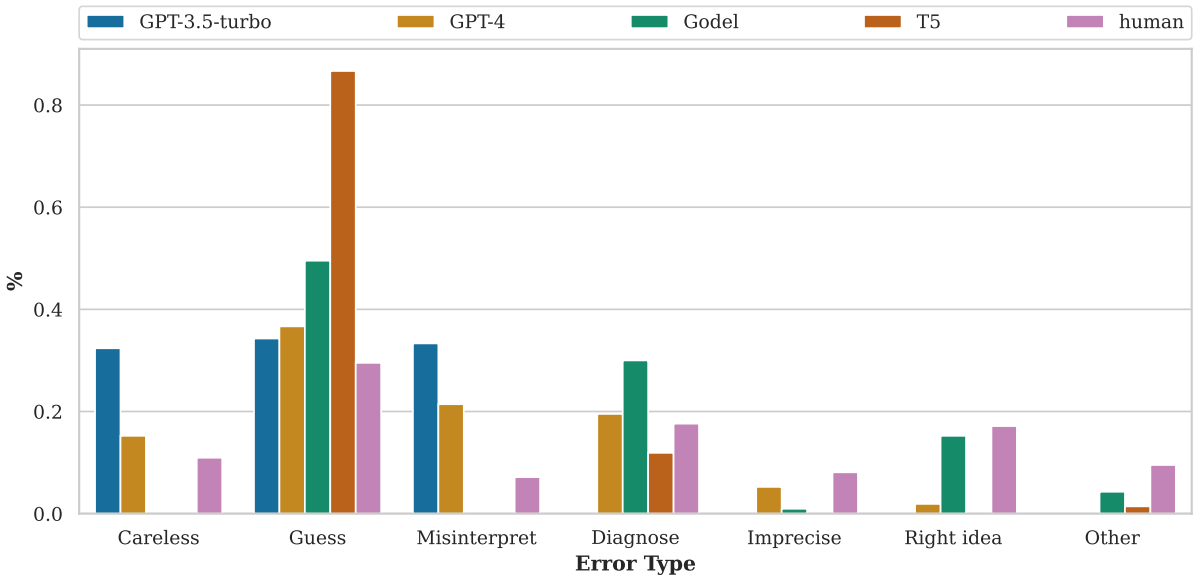
(a)

Method	**Response**
`Original`	That was an incorrect answer.
`Human`	Hmm, what makes you say prime?
`GODEL`	Great try! 39 is a prime number.
`GODEL` (error)	Great try! 39 is a prime number.
`GODEL` (strategy)	Great try! Can you explain how you got 39?
`GODEL` (complete)	Great try! 39 is a prime number. Can you explain how you got 39?

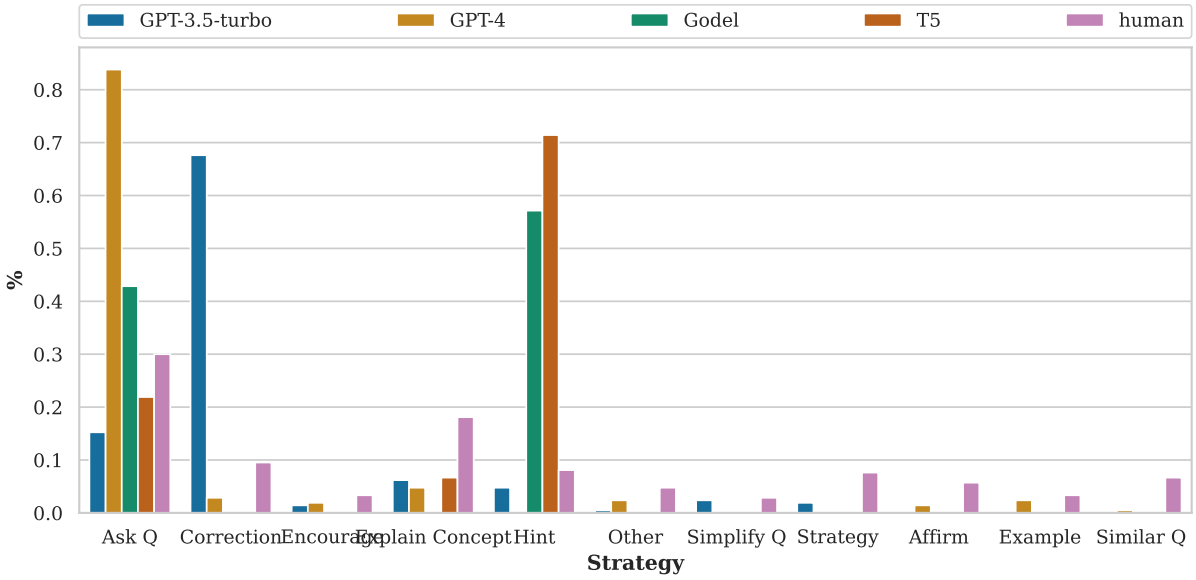
Table 7: (a) show the original context, and (b) shows the responses from the original tutor, teacher and `GODEL` methods. In this example, the error identified is `right-idea`. The strategy is to ask a question, and the intention is to diagnose the student's mistake.

			Automated						**Human evaluation**			
		Method	**ROUGE-1**	**ROUGE-L**	**BLEU**	**Self-BLEU**	**Entropy**	**Length**	**Prefer**	**Useful**	**Care**	**Not Robot**
		`Original`	0.12	0.11	0.01	1.0	3.4	21.53	N/A	N/A	N/A	N/A
		`Human`	0.91	0.91	0.87	0.01	3.44	18.46	**1.26**	**1.19**	**0.86**	0.78
unconstrained	$p(c_r\|c_h)$	`Flan-T5`	0.16	0.15	0.01	0.23	3.9	11.12	0.26	0.25	0.16	0.23
		`GODEL`	0.17	0.16	0.01	0.17	3.75	12.19	0.43	0.28	0.45	0.49
		`GPT-3.5`	0.14	0.12	0.01	0.03	3.28	14.71	0.46	0.45	−0.04	0.22
		`GPT-4`	0.14	0.13	0.01	0.04	2.89	20.2	0.54	0.54	0.50	0.47
error	$p(c_r\|c_h, e)$	`Flan-T5`	0.15	0.14	0.01	0.24	3.67	14.78	−0.07	−0.06	−0.04	−0.01
		`GODEL`	0.17	0.16	0.01	0.16	3.57	13.44	0.36	0.33	0.42	0.38
		`GPT-3.5`	0.13	0.12	0.01	0.03	3.22	13.48	0.41	0.44	0.14	0.17
		`GPT-4`	0.14	0.13	0.01	0.04	2.77	21.66	0.88	0.64	0.79	0.83
strategy	$p(c_r\|c_h, z)$	`Flan-T5`	0.2	0.19	0.02	0.17	3.77	14.26	0.07	−0.01	0.27	0.33
		`GODEL`	0.19	0.19	0.01	0.16	3.63	13.1	0.39	0.16	0.41	0.42
		`GPT-3.5`	0.17	0.15	0.01	0.03	3.21	17.18	0.27	0.29	−0.03	0
		`GPT-4`	0.19	0.17	0.01	0.03	2.85	23.97	0.97	1.08	0.5	**1.07**
complete	$p(c_r\|c_h, e, z)$	`Flan-T5`	0.18	0.18	0.02	0.18	3.79	14.69	0.47	0.36	0.57	0.69
		`GODEL`	0.18	0.17	0.02	0.13	3.62	13.3	0.22	0.20	0.25	0.20
		`GPT-3.5`	0.16	0.14	0.01	0.04	3.11	19.03	0.65	0.58	−0.04	0.59
		`GPT-4`	0.19	0.16	0.02	0.03	2.83	23.8	0.95	0.97	0.7	0.7

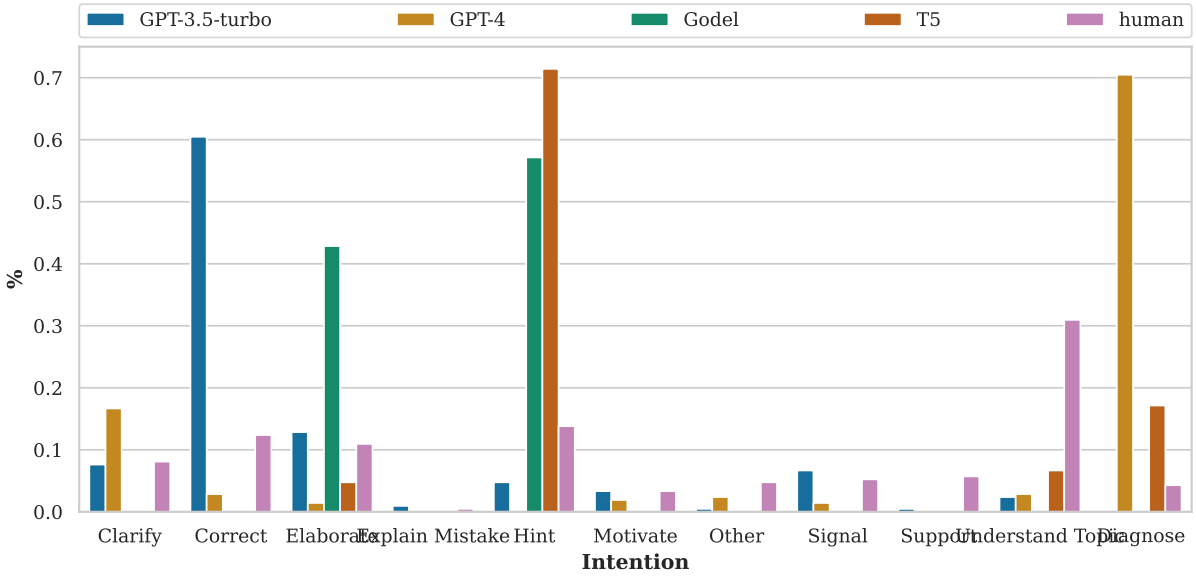
Table 8: Automated metrics and human evaluations. Self-BLEU is reported in bigrams. Entropy is measured with the GPT NeoX 20B model (). Length is the number of space-separated words.



(a) Error Type



(b) Strategy



(c) Intention

Figure 4: Distribution over categories.

Task A prompt

```
You are an experienced elementary math teacher.  Your task is to read a conversation
snippet of a tutoring session between a student and tutor, and determine what type
of error the student makes in the conversation.  We have a list of common errors that
students make in math, which you can pick from.  We also give you the option to write in
your own error type if none of the options apply.

Error list:
0.  Student does not seem to understand or guessed the answer.
1.  Student misinterpreted the question.
2.  Student made a careless mistake.
3.  Student has the right idea, but is not quite there.
4.  Student's answer is not precise enough or the tutor is being too picky about the form
of the student's answer.
5.  None of the above, but I have a different description (please specify in your
reasoning).
6.  Not sure, but I'm going to try to diagnose the student.

Here is the conversation snippet:
Lesson topic:  {lesson_topic}
Conversation:
{c_h}

Why do you think the student made this mistake?  Pick an option number from the error
list and provide the reason behind your choice.  Format your answer as:  [{"answer":  #,
"reason":  "write out your reason for picking # here"}]
```

Figure 5: Prompt used for Task A: Inferring the student's error. `{lesson_topic}` is the placeholder for the lesson topic discussed in the conversation. `{c_h}` is the placeholder for the conversation history leading up to (and including) the student's message that contains the mistake.

Task A prompt [Flan-T5]

```
Determine what type of error the student makes in the conversation.
Lesson topic:  {lesson_topic}
Conversation:
{c_h}
```

Figure 6: Input sequence used for finetuning Flan-T5 for Task A: Inferring the student's error. `{lesson_topic}` is the placeholder for the lesson topic discussed in the conversation. `{c_h}` is the placeholder for the conversation history leading up to (and including) the student's message that contains the mistake.

Task A prompt [GODEL]

```
Instruction: given a conversation snippet of a tutoring session between a student and
tutor and a list of common errors that students make in math, you need to determine
what type of error the student makes in the conversation based on the list. [CONTEXT]
{c_h} [KNOWLEDGE] The problem your student is solving is on topic: {lesson_topic}.
The Error list is guess: Student does not seem to understand or guessed the answer.
misinterpret: Student misinterpreted the question. careless: Student made a careless
mistake. right-idea: Student has the right idea, but is not quite there. imprecise:
Student's answer is not precise enough or the tutor is being too picky about the form of
the student's answer. other: None of the above, but I have a different description
(please specify in your reasoning). diagnose: Not sure, but I'm going to try to
diagnose the student.
```

Figure 7: Input sequence used for finetuning GODEL for Task A: Inferring the student's error. `Instruction` includes basic instructions on the task. `Context` includes `{c_h}`, which is the placeholder for the conversation history leading up to (and including) the student's message that contains the mistake. `{lesson_topic}` is the placeholder for the lesson topic discussed in the conversation, serving as part of the external `Knowledge` in addition to the list of possible student errors mapping labels to descriptions.

Task B prompt

```
You are an experienced elementary math teacher.  Your task is to read a conversation
snippet of a tutoring session between a student and tutor where a student makes a mistake.
You should then determine what strategy you want to use to remediate the student's error,
and state your intention in using that strategy.  We have a list of common strategies and
intentions that teachers use, which you can pick from.  We also give you the option to
write in your own strategy or intention if none of the options apply.

Strategies:
0.  Explain a concept
1.  Ask a question
2.  Provide a hint
3.  Provide a strategy
4.  Provide a worked example
5.  Provide a minor correction
6.  Provide a similar problem
7.  Simplify the question
8.  Affirm the correct answer
9.  Encourage the student
10.  Other (please specify in your reasoning)

Intentions:
0.  Motivate the student
1.  Get the student to elaborate their answer
2.  Correct the student's mistake
3.  Hint at the student's mistake
4.  Clarify a student's misunderstanding
5.  Help the student understand the lesson topic or solution strategy
6.  Diagnose the student's mistake
7.  Support the student in their thinking or problem-solving
8.  Explain the student's mistake (eg.  what is wrong in their answer or why is it
incorrect)
9.  Signal to the student that they have solved or not solved the problem
10.  Other (please specify in your reasoning)

Here is the conversation snippet:
Lesson topic:  {lesson_topic}
Conversation:
{c_h}

How would you remediate the student's error and why?  Pick the option number from the
list of strategies and intentions and provide the reason behind your choices.  Format
your answer as:  [{"strategy":  #, "intention":  #, "reason":  "write out your reason for
picking that strategy and intention"}]
```

Figure 8: Prompt used for Task B: Determining the strategy and intention. `{lesson_topic}` is the placeholder for the lesson topic discussed in the conversation. `{c_h}` is the placeholder for the conversation history leading up to (and including) the student's message that contains the mistake.

Task B prompt [Flan-T5]

```
Determine what strategy you want to use to remediate the student's error, and your
intention in using that strategy.
Lesson topic:  {lesson_topic}
Conversation:
{c_h}
```

Figure 9: Input sequence used for finetuning Flan-T5 for Task B: Determining the strategy and intention. `{lesson_topic}` is the placeholder for the lesson topic discussed in the conversation. `{c_h}` is the placeholder for the conversation history leading up to (and including) the student's message that contains the mistake.

Task B prompt [GODEL]

```
Instruction: given a conversation snippet of a tutoring session between a student
and tutor and a list of common strategies and intentions that teachers use, you need
to determine what strategy you want to use to remediate the student's error, and your
intention in using that strategy. [CONTEXT] {c_h} [KNOWLEDGE] The problem your student
is solving is on topic: {lesson_topic}. Strategies: explain_concept: Explain a
concept; ask_question: Ask a question; provide_hint: Provide a hint; provide_strategy:
Provide a strategy; provide_example: Provide a worked example; provide_correction:
Provide a minor correction; provide_similar_problem: Provide a similar problem;
simplify_question: Simplify the question; affirm_correct_answer: Affirm the correct
answer; encourage_student: Encourage the student; other: Other (please specify in your
reasoning). Intentions: motivate_student: Motivate the student; elaborate_answer:
Get the student to elaborate their answer; correct_mistake: Correct the student's
mistake; hint_mistake: Hint at the student's mistake; clarify_misunderstanding:
Clarify a student's misunderstanding; understand_topic: Help the student understand
the lesson topic or solution strategy; diagnose_mistake: Diagnose the student's
mistake; support_thinking: Support the student in their thinking or problem-solving;
explain_mistake: Explain the student's mistake (eg. what is wrong in their answer or
why is it incorrect); signal_goal: Signal to the student that they have solved or not
solved the problem; other: Other (please specify in your reasoning).
```

Figure 10: Input sequence used for finetuning GODEL for Task Task B: Determining the strategy and intention. `Instruction` includes basic instructions on the task. `Context` includes `{c_h}`, which is the placeholder for the conversation history leading up to (and including) the student's message that contains the mistake. `Knowledge` refers to the external environment or knowledge, which includes `{lesson_topic}`, the list of possible strategies, and the list of possible intentions.

Task C prompt

```
You are an experienced elementary math teacher and you are going to respond to a
student's mistake in a useful and caring way. The problem your student is solving is
on topic: {lesson_topic}.{e}{z}
{c_h}
tutor (maximum one sentence):
```

Figure 11: Prompt used for Task C: Generating a remediation response. `{lesson_topic}` is the placeholder for the lesson topic discussed in the conversation. If the error is shown (eg. in the *error-constrained* ablation), `{e}` is a placeholder for the error type identified by the educator—note that `{e}` would be formatted to be a coherent piece of text. If the strategy and intention are shown (eg. in the *strategy-constrained* ablation), `{z}` is a placeholder for the strategy and intention picked by the educator—note that, similarly, `{z}` would be formatted to be a coherent piece of text. `{c_h}` is the placeholder for the conversation history leading up to (and including) the student's message that contains the mistake. We add an additional constraint "(maximum one sentence)" because from our experiments, `gpt-3.5-turbo` and `gpt-4` typically output extremely long responses that would be unnatural for this tutoring conversation domain.

Task C prompt [Flan-T5]

```
You are an experienced elementary math teacher and you are going to respond to a
student's mistake in a useful and caring way. The problem your student is solving is
on topic: {lesson_topic}.{e}{z}
{c_h}
tutor:
```

Figure 12: Input sequence used for finetuning Flan-T5 for Task C: Generating a remediation response. For descriptions of placeholders - `{lesson_topic}`, `{e}`, `{z}` and `{c_h}`, reference Figure 11 for details.

Task C prompt [GODEL]

```
You are an experienced elementary math teacher and you are going to respond to a
student's mistake in a useful and caring way.{z} [CONTEXT] {c_h} [KNOWLEDGE] The problem
your student is solving is on topic {lesson_topic}.{e}
```

Figure 13: Input sequence used for finetuning GODEL for Task C: Generating a remediation response. `{z}`, if shown, is formatted to be part of the `Instruction`; `{c_h}` is formatted to be the dialogue history `Context`; `{lesson_topic}` and `{e}`, if shown, are formatted to be part of the `Knowledge`. For descriptions of placeholders - `{lesson_topic}`, `{e}`, `{z}` and `{c_h}`, reference Figure 11 for details.

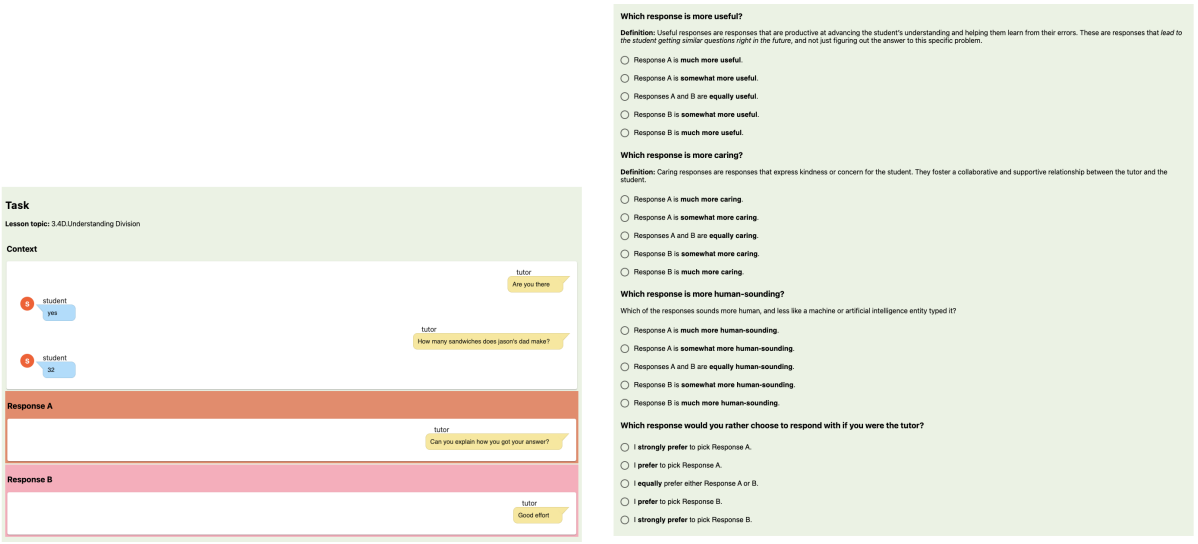
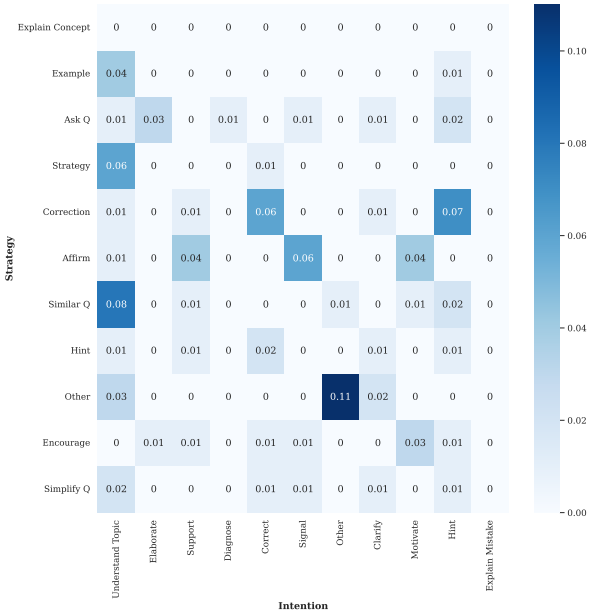
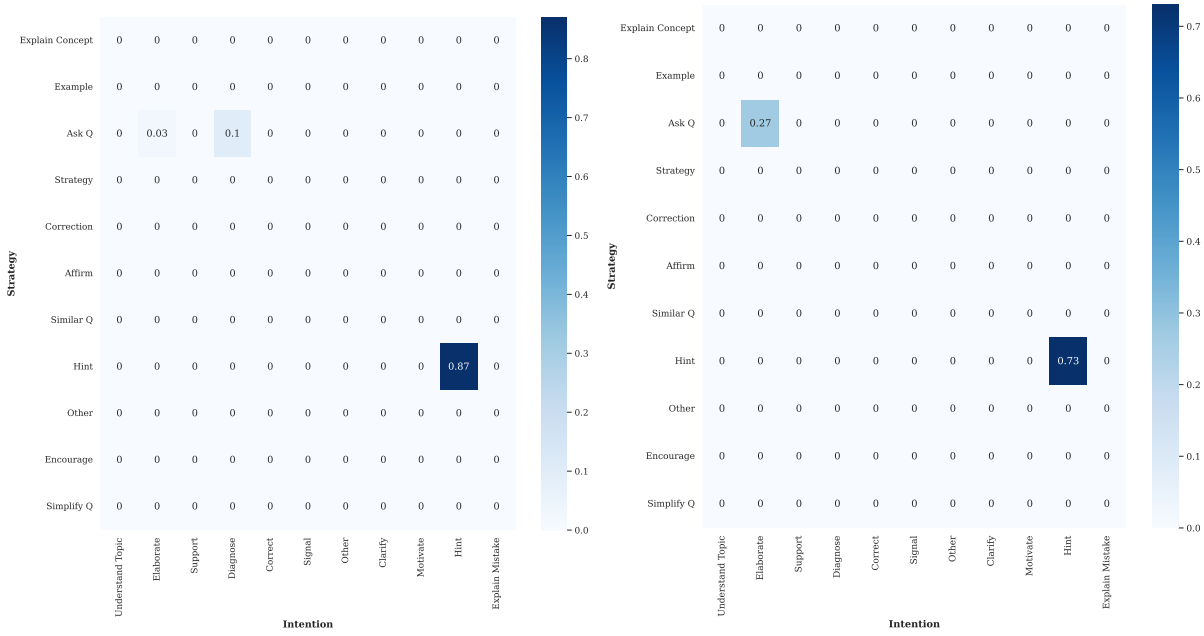


Figure 14: Annotation interface for evaluating Task C remediation responses.

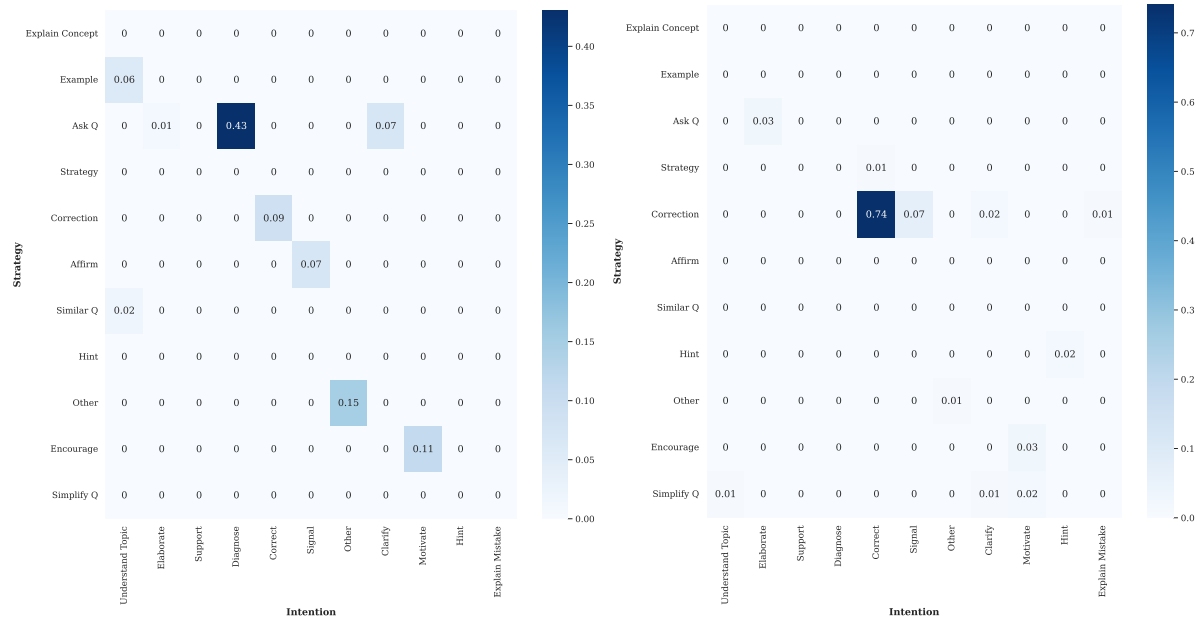


(a) Human



(b) Flan-T5

(c) GODEL



(d) gpt-4

(e) gpt-3.5-turbo

Figure 15: Comparing pairs of strategies and intentions.

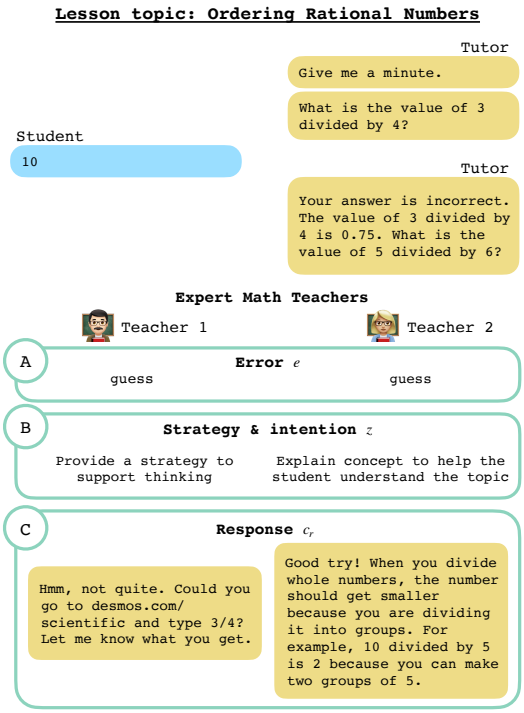

Lesson topic: Ordering Rational Numbers
Tutor
Give me a minute.
What is the value of 3 divided by 4?
Student
10
Tutor
Your answer is incorrect. The value of 3 divided by 4 is 0.75. What is the value of 5 divided by 6?
Expert Math Teachers
Teacher 1
Teacher 2
A
Error e
guess
guess
B
Strategy & intention z
Provide a strategy to support thinking
Explain concept to help the student understand the topic
C
Response c_r
Hmm, not quite. Could you go to desmos.com/scientific and type 3/4? Let me know what you get.
Good try! When you divide whole numbers, the number should get smaller because you are dividing it into groups. For example, 10 divided by 5 is 2 because you can make two groups of 5.

Figure 16