Bridging the Novice-Expert Gap via Models of Decision-Making: A Case Study on Remediating Math Mistakes

Anonymous ACL submission

Abstract

Scaling high-quality tutoring remains a major challenge in education. Due to growing demand, many platforms employ novice tutors who, unlike experienced educators, struggle to address student mistakes and thus fail to seize prime learning opportunities. Our work explores the potential of large language models (LLMs) to close the novice-expert knowledge gap in remediating math mistakes. We contribute `Bridge`, a method that uses cognitive task analysis to translate an expert's latent thought process into a decision-making model for remediation. This involves an expert identifying (A) the student's error, (B) a remediation strategy, and (C) their intention before generating a response. We construct a dataset of 700 real tutoring conversations, annotated by experts with their decisions. We evaluate state-of-the-art LLMs on our dataset and find that the expert's decision-making model is critical for LLMs to close the gap: responses from GPT4 with expert decisions (e.g., "simplify the problem") are +76% more preferred than without. Additionally, context-sensitive decisions are critical to closing pedagogical gaps: random decisions decrease GPT4's response quality by -97% than expert decisions. Our work shows the potential of embedding expert thought processes in LLM generations to enhance their capability to bridge novice-expert knowledge gaps.

1 Introduction

Human tutoring plays a critical role in accelerating student learning, and is one of the primary ways to combat 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 novices with proper training can be effective tutors (Nickow et al., 2020).

Responding to student mistakes in real-time is a critical area where novice tutors tend to struggle. Mistakes are prime learning opportunities to address misconceptions (Boaler, 2013), but effective responses involve pedagogical expertise in engaging with student's thinking and building positive rapport (Roorda et al., 2011; Pianta, 2016; Shaughnessy et al., 2021; Robinson, 2022). Novices typically learn from experts to understand the expert's thought process however hiring experienced educators to provide timely feedback is resource-intensive (Kraft et al., 2018; Kelly et al., 2020).

One potential solution is the use of automated tutors (Graesser et al., 2004). With recent advances in large language models (LLMs), this approach has gained even more interest (Khan Academy, 2023). However their ability to remediate is yet to be evaluated. Prior work suggests several shortcomings with LLMs, including lacking reliable subject and pedagogical knowledge (Frieder et al., 2023; Wang and Demszky, 2023; Singer, 2023), that can be mitigated using explicitly thought processes such as through chain-of-thought prompting (Wei et al., 2022).

To address these challenges, our work makes several key contributions. First, we build **`Bridge`, a method that leverages cognitive task analysis to elicit the latent thought processes of experts**. We apply `Bridge` to remediation where we collaborate extensively with experienced math educators to translate their thought process into a decision-making model. `Bridge` breaks down the experts' thought process: illustrated in Figure 1, Step A is to infer the student's error (e.g., the student guessed); Step B is to determine the remediation strategy (e.g., provide a solution approach); and Step C is to identify the strategy intention (e.g., to help the student understand the concept).

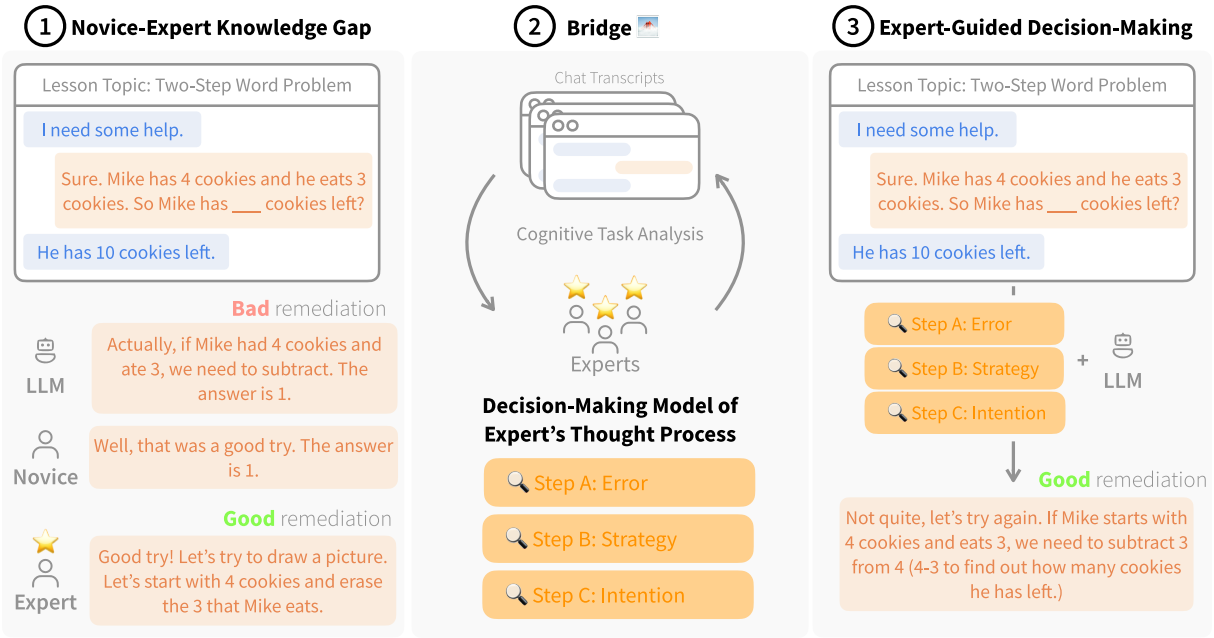


Figure 1: ① **Closing the knowledge gap at scale.** LLMs and novice tutors lack the pedagogical knowledge to engage with student mistakes, yet they are readily available for 1:1 tutoring. Experts like experienced teachers have the pedagogical knowledge, but are hard to scale. ② **How do we model the expert's thought process?** Our work builds `Bridge` which leverage cognitive task analysis to translate the latent thought process of experts into a decision-making model. ③ **Applying `Bridge` with LLMs.** To bridge the knowledge gap, we scale the expert's knowledge with LLMs using the expert-guided decision-making model.

We construct **a dataset of real-world tutoring conversations, annotated with expert decisions and responses.** Our open-source dataset consists of 700 real tutoring sessions conducted with 1st-5th grade students in Title I schools, predominantly serving low-income students of color. Following FERPA guidelines, our study is IRB-approved and conducts secondary data analysis based on our Data Use Agreement with the tutoring provider and school district.

We conduct a **thorough human evaluation to compare the expert, novice and LLMs in remediation**. To our knowledge, our work is the first to assess the performance of LLMs such as GPT4 and instruct-tuned Llama-2-70b on remediating student mistakes. We find that the response quality of LLMs significantly improve with the expert's decision-making process: Response from GPT4 with expert- and self-generated decisions are 76-88% more preferred than GPT4 without. Context-sensitive decisions are also critical to closing the knowledge gap: Random decisions decrease GPT4's response quality -67% than expert decisions. Complementing our quantitative analysis, our **lexical analysis reveals that novices and LLMs without the expert's decision-making process engage superficially with student's problem-solving process**: They give away the answer or prompt the student to re-attempt without further guidance ("double check", "try again").

2 Related Work

2.1 Modeling the Decision-Making Process of Experts

Cognitive task analysis (CTA) uncovers the latent decision-making process of experts across a range of domains such as education, medicine and law (Ryder and Redding, 1993; Clark et al., 2008; Klein, 2015). CTA decode the *observable actions* (e.g., the expert's remediation responses) into the *latent mental processes* that generate the observable actions (e.g., the expert's inferences about the student's mistake). A key application area of CTA is to close knowledge gaps through real-time decision aids that enhance the cognitive skills of novices (Hall et al., 1995; Gagne and Medsker, 1996; Van Merriënboer, 1997; Klein, 2008; Zsambok and Klein, 2014); Lee (2004) discusses the significant improvements in novices with CTA across multiple disciplines. While previous NLP works have developed methods for auto-labeling CTA transcripts (Du et al., 2019), less work has been

done on synthesizing models of expert decision processes for natural language generation or contributing data with expert decisions. Our work contributes both the `Bridge` method and an accompanying dataset to this end.

2.2 Responding to Student Mistakes in Mathematics

Recognizing misconceptions is key to facilitating meaningful student learning and retention (Stefanich and Rokusek, 1992; Wilcox and Zielinski, 1997; Riccomini, 2005; Stein et al., 2005; Schnepper and McCoy, 2013). Effective remediation coincides with educators engaging with the mathematical details in student responses, which in turn fosters strong teacher-student relationships and student motivation (Wentzel, 1997; Pianta et al., 2003; Robinson, 2022; Wentzel, 2022; Easley and Zwoyer, 1975; Brown and Burton, 1978; Carpenter et al., 1999, 2003; Lester, 2007; Loewenberg Ball and Forzani, 2009). Prior education research discusses multiple good practices in remediating student mistakes, ranging from visual aids (CAST, 2018) to the Socratic method (Lepper and Woolverton, 2002). However, less work has been done to understand the thought process of an experienced educator of when, how and why they use one strategy over another.

2.3 Automated Feedback in Education

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; Jacobs et al., 2022; Demszky and Liu, 2023; Wang and Demszky, 2023; Demszky et al., 2023). 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 and hallucinate (Frieder et al., 2023; Ji et al., 2023). A human tutor in-the-loop is key in catching these undesirable responses. Our work is related to human-LLM approaches that leverage expert-informed linguistic attributes (Sharma et al., 2023; Handa et al., 2023). However, critically, our work is about modeling *the expert's latent thought process* behind their responses, such as their strategy choices and intentions, rather than the *observable* linguistic attributes. We explore the potential of leveraging expert-informed decision-making processes for bridging knowledge gaps and constructing human-LLM interaction frameworks grounded in expertise.

3 Data Sources

Tutoring transcripts. Our data is sourced from a tutoring provider that offers end-to-end services for school districts, including the tutoring platform, instructional materials, and tutors. The research team executed Data Use Agreements with the tutoring provider and Southern U.S. school district serving over 30k that outlined the allowable usage of the data to improve instruction in collaboration with an educational agency. Following FERPA guidelines, we were eligible to engage in secondary data analysis with student data, which is what we did for this study. The students in these tutoring sessions are in the first to fifth grade, learning a variety of math topics. The majority of schools are classified as Title I and three-quarters of students identify as Hispanic/Latinx. This district focused on addressing existing achievement gaps among their students, as well as responding to the learning disruptions caused by the pandemic. The tutoring interactions are text-based, integrated on the providers' online platform. The platform has several features, including a whiteboard. 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]. Our data 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.

4 The Bridge Method for Expert-Guided Decision-Making

We introduce Bridge which uses cognitive task analysis (CTA) to analyze the experts' latent thought process (§4.1). We translate it into a decision-making process (§4.3), where each step is associated with a set of decision options (§4.2).

4.1 Cognitive Task Analysis

We conduct CTA with four experienced math teachers to develop a model of their decision-making process for remediation.

Collaboration with experts. We collaborated extensively with math teachers, spanning across several months. 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 compensate the teachers developing the decision-making framework $50/hour. We compensate the teachers annotating the dataset with their decision steps and responses at $40/hour.

Our objective is to faithfully capture their step-by-step decision process and develop a comprehensive set of decision options for each step. We work with two math teachers to develop the decision-making process for remediation, and validated it with two other math teachers. We conduct CTA through a series of observations and interviews, which involved cataloging patterns in their decisions; Cooke (1999) provides a comprehensive overview of other CTA methods.

Development of decision-making process. We provide the experts conversation examples containing student mistakes (identified from §3) and asked them to directly revise the tutor's remediation response to be more useful and caring. The experts and co-author met on a weekly basis where we went through the experts' revisions and discussed their approaches to each mistake. We used three questions to facilitate the discussion: (1) *What* did the experts notice? (2) *How* did they want to react? and (3) *Why* did they want to react in that way? Themes emerged after a few meetings. Based on their own experiences, experts inferred the student's level of understanding as context for their remediation response. This resulted in *Step A: Infer the student's error* to answer the first question. Experts used several techniques to engage with the student's error, such as asking questions and simplifying the problem to meet the student's level of understanding. The diverse strategies led to *Step B: Determine the strategy*. Finally, the experts used strategies for different ends depending on error. For example, they might ask a question to hint at the mistake or diagnose the student. This insight resulted in *Step C: Identify the intention behind the strategy*. We verified that this decision-making model mimicked their thought process by asking them to apply it to new tutoring conversations. We additionally verified it with two other experts who could seamlessly use it during their remediation. For additional information about the development process, please refer to Appendix A.

Development of decision options. We created decision options for each step and edited the options through more iterations of the experts remediating using the step-by-step decision-making process. The options were finalized once the experts and the co-authors were satisfied with the coverage and with the natural fit of the model to the teachers' remediation process.

4.2 Decision Options

This section details each step's decision options. Due to space reasons, please refer to Appendix B for examples of each option.

4.2.1 Step A: Infer the Type of Error

Identifying the student's error is prerequisite to successful remediation (Easley and Zwoyer, 1975; Bamberger et al., 2010). Our approach intends to support novices who are not necessarily content experts. Therefore we define "error" as a student's degree of understanding, which aligns with literature on math curriculum design and psychometrics that maintain continuous scales of student understanding (Gagne, 1962, 1968; White, 1973; Resnick et al., 1973; Glaser and Nitko, 1970; Vygotsky and Cole, 1978; Wertsch, 1985; Embretson and Reise, 2013). As such, our error categories are topic-agnostic descriptions of a student's understanding, and complement the topic-agnostic strategies in Step B. The categories are: `guess`: The student does not seem to understand or guessed the answer; `misinterpret`: The student misinterpreted the question; `careless`: The student made

a careless mistake; `right-idea`: The student has the right idea, but is not quite there[1]; `imprecise`: The student's answer is not precise enough or the tutor is being too picky about the form of the student's answer; `not-sure`: Not sure, but I'm going to try to diagnose the student (used sparingly); `N/A`: None of the above (used sparingly).

4.2.2 Step B: Determine the Strategy

Errors are persistent unless the teacher intervenes pedagogically with a strategy that guides the student's understanding (Radatz, 1980). The strategies are: 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.

4.2.3 Step C: Identify the Intention

The intentions are: Motivate the student, Get the student to elaborate their answer, Correct the mistake, Hint at the mistake, Clarify the misunderstanding, Help the student understand the lesson topic or solution strategy, Diagnose the mistake, Support the student in their thinking or problem-solving, Explain the 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.

4.3 Formalism for Expert Decision-Making Process in Remediation

Given a conversation history c_h, we formalize the expert's responses c_r^* as being generated from the following computational model:

$$c_r^* \sim p(c_r|c_h, \underbrace{e}_{\text{Step A}}, \underbrace{z_{\text{what}}}_{\text{Step B}}, \underbrace{z_{\text{why}}}_{\text{Step C}}),$$

where e is the error, z_{what} the strategy, and z_{why} the intention. Our dataset contains 700 examples, where each example is $(c_h, c_r', e, z_{\text{what}}, z_{\text{why}}, c_r^*)$. Each example contains the conversation history c_h which includes the lesson topic and the last 5 conversation messages leading up to the student's turn where the mistake is made; i.e., $c_h[-1]$ is the student's conversation turn where they make a mistake. It also contains the novice tutor's original response to the student's mistake c_r' and the experts' decision annotations and responses. We split the final dataset into a train, validation, and test set with a 6:1:3 ratio. The train set contains 420, validation 70, and test 210 examples.

[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.

5 Experiments

5.1 Models

We compare the expert-written responses against three state-of-the-art models `gpt-4`, `gpt-3.5-turbo`, and `llama-2-70b-chat` (Touvron et al., 2023) in a 0-shot setting on the test set. During our preliminary experiments, we also evaluated Falcon-40b-Instruct (Almazrouei et al., 2023), Flan-T5 (large) (Chung et al., 2022), the goal-directed dialog model GODEL (large) (Peng et al., 2022) zero-shot and few-shot. We also finetuned Flan-T5 and GODEL. However, we found the models' responses to be very poor upon manual inspection or evaluated as much worse in human evaluations than the other three models. Therefore, we have omitted their results from the paper. We use greedy decoding for all models.

5.2 Task Setup

We evaluate the model responses under different decision-making conditions. The model prompts are in Appendix D; each prompt includes instructions to respond in a useful and caring way.

1. *No decision-making*: Models directly respond, $c_r \sim p(c_r|c_h)$. This condition is compared against models with the `Bridge` decision-making framework.
2. *Expert decision-making*: Models generate with the expert's decisions, $c_r \sim p(c_r|c_h, e, z_{\text{what}}, z_{\text{why}})$.
3. *Self decision-making*: Models make their own decisions, then generate responses based on them, $c_r \sim p(c_r|c_h, e^{\text{model}}, z_{\text{what}}^{\text{model}}, z_{\text{why}}^{\text{model}})$. We compare the models' decisions to the experts' as well as the impact of the decisions on the response quality.
4. *Random decision-making*: We randomly select decisions. We can determine the importance of context-sensitive decisions with this condition.

Method		Prefer	Useful	Care	Not Robot	Overall
Condition	Model c_r					
	Expert	**1.26**	**1.19**	**0.86**	**0.78**	**1.02**
None	Llama-2	0.49	0.48	0.45	0.68	0.53
None	GPT-3.5	0.47	0.47	−0.04	0.23	0.28
None	GPT-4	0.54	0.54	0.50	0.47	0.51
Expert	Llama-2	0.61	0.56	0.37	0.41	0.49
Expert	GPT-3.5	0.65	0.58	−0.04	0.59	0.45
Expert	GPT-4	**0.95**	**0.97**	**0.70**	**0.70**	**0.83**
Self	Llama-2	0.91	0.97	0.29	0.62	0.70
Self	GPT-3.5	0.36	0.33	−0.17	0.15	0.16
Self	GPT-4	**1.02**	**1.05**	**0.62**	**0.68**	**0.84**
Random	Llama-2	0.35	0.32	0.15	0.60	0.35
Random	GPT-3.5	0.20	0.12	0.10	0.28	0.17
Random	GPT-4	0.32	0.36	−0.13	0.51	0.26

Table 1: **Human evaluations.** The expert-written responses are grayed as a reference. The highest column values are **bolded**. Highest values amongst LLMs are highlighted. Two rows are highlighted if they are not statistically different from each other with a two-sided t-test.

6 Evaluation

6.1 Human evaluation of response quality.

We measure the extent to which the generated responses improve over the original tutors' responses. We recruit teachers through Prolific (identified through Prolific's screening criteria) to perform pairwise comparisons between the tutor response and a response generated by the expert or one of the 12 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 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 is *human-soundingness*; our preliminary analysis indicated that low learning outcomes strongly correlated with whether the student was distracted by whether their tutor was human during their tutoring session. Given that the tutoring is chat-based, we include this as another dimension for measuring effectiveness. Finally, we ask the raters which responses they *prefer* using, if they were the tutor. Each dimension is rated on a 5-point Likert scale. We convert the ratings to integers between -2 and 2: -2 indicates the rater much more prefers the original tutor's response and 2 for the alternative response. Please refer to Appendix E for more information on the human evaluation setup.

6.2 Lexical analysis and qualitative examples.

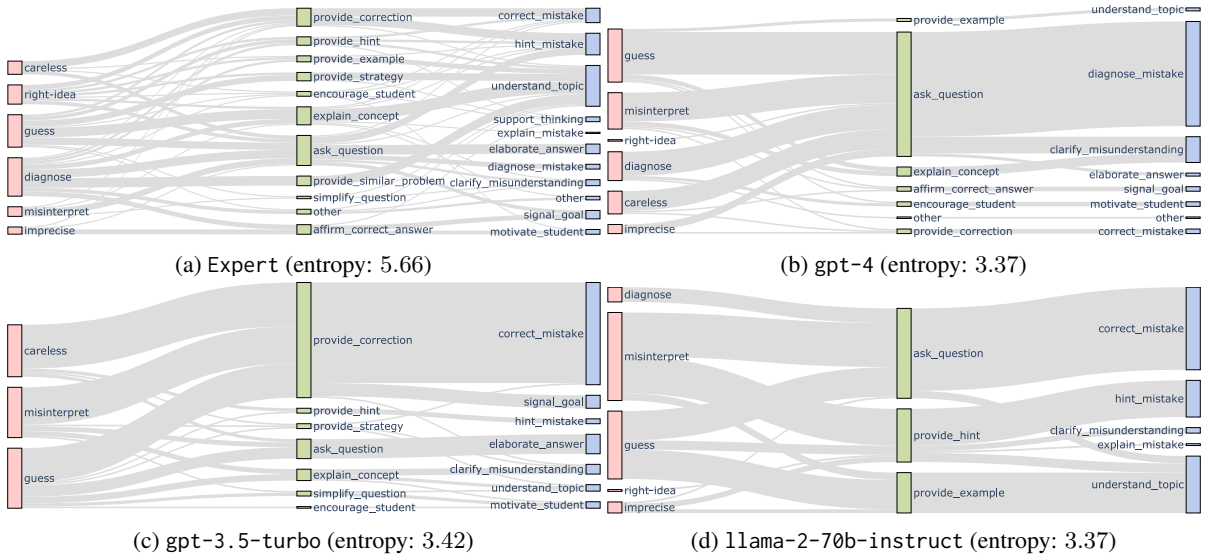
We perform a lexical analysis to understand the linguistic differences caused by the expert's decision-making model. We compute the log odds ratio, latent Dirichlet prior, measure defined in Monroe et al. (2008) to estimate the distinctiveness of a bigram appearing in a response source. We consider the response sources to be from GPT4 in all four decision-making conditions listed in Section 5.2; please refer to Appendix F for additional lexical analysis. We pre-process the data using Python's NLTK package for tokenization and lowercasing, and discard stop words and non-alphanumeric tokens (Bird et al., 2009). We use the Gensim Phrases Python package to retrieve frequent bigrams in the dataset (Rehurek and Sojka, 2011).

7 Results

7.1 Human evaluations of response quality.

Table 1 summarizes the results. Notably, there is a **large gap between the experts and models in the *no decision-making* condition** (up to 2.6x better overall). Even though models in the *no decision-making* condition consistently outperform the original tutor responses (indicated by the positive values) on most dimensions, the gap in response quality may indicate the pedagogical knowledge gap between experts and LLMs.

We observe that **the *expert decision-making* condition outperforms the *no decision-making* condition**, particularly on "prefer" (+76% on



(a) `Expert` (entropy: 5.66)

(b) `gpt-4` (entropy: 3.37)

(c) `gpt-3.5-turbo` (entropy: 3.42)

(d) `llama-2-70b-instruct` (entropy: 3.37)

Figure 2: **Expert decision-making paths are diverse whereas LLM's are less diverse.** The entropy of decision paths is shown in the subcaption: The experts' paths have higher entropy and thus are more diverse than those of the LLMs. The red left column is Step A's error decision; green middle column is Step B's strategy decision; and blue right column is Step C's intention decision.

`gpt-4`) and "useful" (+80% on `gpt-4`). The improvement in overall score is statistically significant for all models under a two-tailed t-test ($p < 0.05$). Surprisingly, the *expert decision-making* condition for `llama-2` and `gpt-3.5-turbo` does not improve on "care". We attribute this to the challenges in generating responses that are both technically instructive ("useful") and emotionally supportive ("care") for the student.

How well can models self-improve by selecting their own decisions? **`llama-2` and `gpt-4` in the *self decision-making* condition significantly outperform their *no decision-making* counterparts on "prefer" and "useful"** ($p < 0.05$, up to +95%). However, this is not the case for `gpt-3.5-turbo` with *self decision-making*. We hypothesize this is due to its poor decisions and confirm this in Figure 2. Figure 2 illustrates the decision paths from the experts and the LLMs in *self decision-making* on the test examples and reports the path entropy. The width is the proportion of error types that is subsequently treated with which strategy and with which intention. `gpt-3.5-turbo` overwhelmingly corrects the student's mistake whereas the other models rely on other strategies. This suggests that **directly correcting the student's mistake is not always a good decision and that poor decisions reinforce poor response quality**.

Figure 2 reveals another interesting observation: **Experts exhibit diverse decision paths, whereas LLMs do not.** Our work provides additional evidence of homogenization effects in LLMs (Padmakumar and He, 2023). This prompts another question: Does deliberate decision-making matter, or could we randomly pick decisions to encourage similar diversity? Deliberate decisions do matter: **Models with *random decision-making* perform significantly worse than their *expert decision-making* condition on the "overall" score** ($p < 0.05$), sometimes even worse than models with *no decision-making* ($p < 0.05$ for `gpt-4`, `llama-2`).

7.2 Lexical Analysis

Table 2 highlights the differences in word usage across the `GPT4` decision-making conditions, and Table 3 shows an example of the word usage in context. Table 2 suggests that the high human evaluations for `GPT4` with *expert* or *self decision-making* are because they engage more with the problem-solving process (e.g.,, "explain_steps"). The lowly evaluated settings—`GPT4` with *no* or *random decision-making*—weakly engage with the problem-solving process, only acknowledging the student's effort (e.g., "appreciate_effort" in Table 2) or even giving away the answer (e.g.,, "Actually, the correct answer is 9" in Table 3). Altogether, these results suggest that the effective use of the decision-making model guides LLMs to support the student's problem-solving process, rather than engage superficially with the student's final answer.

None + GPT4		Expert + GPT4		GPT4 + GPT4		Random +GPT4	
bigram	log odds	bigram	log odds	bigram	log odds	bigram	log odds
lets_closer	2.42	steps_took	1.93	explain_arrived	3.71	start_remember	1.73
closer_look	2.42	thought_process	1.54	can_explain	3.59	good_try	1.69
appreciate_effort	2.23	review_concept	1.54	arrived_answer	2.99	thats_good	1.59
effort_lets	2.23	took_arrive	1.43	are_sure	2.24	thats_right	1.59
look_problem	2.01	explain_steps	1.43	sure_that	2.24	try_again	1.39
correct_solution	1.86	help_understand	1.30	arrived_number	1.73	good_start	1.37
sure_weve	1.61	understand_concept	1.30	and_long	1.41	lets_break	1.24
good_attempt	1.61	number_sides	1.30	answer_step	1.41	example_apples	1.22
lets_steps	1.61	lets_step	1.26	numbers_sure	1.41	represents_glasses	1.22
great_effort	1.33	explain_thought	1.16	can_tell	1.33	original_number	1.22

Table 2: Top 10 bigrams. **GPT4 with *expert-* or *self decision-making* engages more with the student's problem-solving process**, by asking the student for explanations and their reasoning process. GPT4 with *no and random decision-making* engages superficially with the student's answer, by acknowledging their attempt or prompting them to try again without further guidance.

Lesson Topic	Addition and Subtraction Within 100
Context (c_h)	Novice: We need to subtract 6 from 15. Student: oh okay... Novice: What is is value of 15 - 6? Student: it is 11?
Expert	Ah, almost! Here's how we can know for sure. We start at 15 and we need to go down 6. So let's count down 6 numbers from 15. 14, 13, 12, 11, 10, 9.
Novice	That was a good effort.
GPT4 (None)	That's correct! Great job, 15 - 6 equals 11. Keep up the good work!
GPT4 (Expert)	That's close, but let's try again. If we break down 15 into 10 and 5, we can subtract 6 from 10 first, which gives us 4. Then, add the 5 back, so 4 + 5 = 9. So, 15 - 6 = 9.
GPT4 (Self)	Are you sure about that? Let's try subtracting 6 from 15 again.
GPT4 (Random)	Actually, the correct answer is 9.

Table 3: Response example. **The Expert and GPT4 with *expert decision-making* engage with the problem-solving process, whereas the other settings engage with the student's answer**, such as giving away the answer or acknowledging the student's attempt. In this example, GPT4 (None) also responds with the incorrect answer.

8 Discussion & Conclusion

Our work presents several contributions for bridging the expert-novice gap and improving the learning experience at scale. First, we develop Bridge, which leverages cognitive task analysis to translate an expert's latent thought process into a decision-making framework. We apply this to the task of remediating mistakes because they are prime learning opportunities to correct misunderstandings hindering learning. Second, we contribute a rich dataset with expert annotations on their decisions and responses. The dataset comes from a tutoring program that works with a majority of Title I schools, and is a valuable resource for providing equitable, high-quality learning experiences. Finally, we perform a thorough evaluation and lexical analysis of experts, novices and LLMs. We demonstrate that expert-guided decision-making and strategic decision selection are critical to improving remediation quality. Novices and LLMs alone use passive remediation language and do not engage with the student's error traces. Our findings indicate promising avenues for scaling high-quality tutoring with expert-guided decision-making. For example, the tutor can make the decisions and the LLM generates an initial response that is further edited by tutor. Altogether, our work shows promising results of an expert-guided human-LLM approach that makes strides towards bridging the knowledge gap.

9 Limitations and Future Work

While our work provides a useful starting point for leveraging expert decision-making models at scale and remediating student mistakes, there are limitations to our work. Addressing these limitations will be an important area for future research.

Collapsing expert thought processes. LLMs and novices might still receive incomplete information or maintain misconceptions when following the expert's decision-making process, because the process distills the expert's knowledge. Nonetheless, we hope `Bridge` and the accompanying dataset provide a useful foundation for leveraging expert knowledge at scale.

Experts. We work with a handful of experts based on the U.S., which is not representative of experienced teaching backgrounds from other countries or cultures. We hope that future work can build on `Bridge` and adapt the decision-making models to fit to other expert pools.

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.

Expanding to other subjects. Our dataset and benchmark currently focuses on mathematics. The remediation process for mathematics and the decision options may not directly transfer to other subjects, although they 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 for our 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].

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

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

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

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.

Richard E Clark, David F Feldon, Jeroen JG Van Merrienboer, Kenneth A Yates, and Sean Early. 2008. Cognitive task analysis. In *Handbook of research on educational communications and technology*, pages 577–593. Routledge.

Nancy J Cooke. 1999. Knowledge elicitation. *Handbook of applied cognition*, pages 479–509.

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.

Junyi Du, He Jiang, Jiaming Shen, and Xiang Ren. 2019. Eliciting knowledge from experts: Automatic transcript parsing for cognitive task analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4280–4291.

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

Susan E Embretson and Steven P Reise. 2013. *Item response theory*. Psychology Press.

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.

Robert M Gagne. 1962. The acquisition of knowledge. *Psychological review*, 69(4):355.

Robert M Gagne. 1968. Presidential address of division 15 learning hierarchies. *Educational psychologist*, 6(1):1–9.

Robert M Gagne and Karen L Medsker. 1996. The conditions of learning: Training applications.

Robert Glaser and Anthony J Nitko. 1970. Measurement in learning and instruction.

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.

Ellen P Hall, Sherrie P Gott, and Robert Alan Pokorny. 1995. *A procedural guide to cognitive task analysis: The PARI methodology*. Armstrong Laboratory, Air Force Materiel Command.

Kunal Handa, Margaret Clapper, Jessica Boyle, Rose E Wang, Diyi Yang, David S Yeager, and Dorottya Demszky. 2023. "mistakes help us grow": Facilitating and evaluating growth mindset supportive language in classrooms.

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].

Gary Klein. 2008. Naturalistic decision making. *Human factors*, 50(3):456–460.

Gary Klein. 2015. A naturalistic decision making perspective on studying intuitive decision making. *Journal of applied research in memory and cognition*, 4(3):164–168.

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.

Robin Louise Lee. 2004. *The impact of cognitive task analysis on performance: A meta-analysis of comparative studies*. University of Southern California.

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].

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

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.

Vishakh Padmakumar and He He. 2023. Does writing with language models reduce content diversity? *arXiv preprint arXiv:2309.05196*.

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.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2):2.

Lauren B Resnick, Margaret C Wang, and Jerome Kaplan. 1973. Task analysis in curriculum design: A hierarchically sequenced introductory mathematics curriculum 1. *Journal of Applied Behavior Analysis*, 6(4):679–709.

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.

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

Joan M Ryder and Richard E Redding. 1993. Integrating cognitive task analysis into instructional systems development. *Educational Technology Research and Development*, 41(2):75–96.

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.

Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9977–10000, Toronto, Canada. Association for Computational Linguistics.

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.

Natasha Singer. 2023. In classrooms, teachers put a.i. tutoring bots to the test.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jeroen JG Van Merriënboer. 1997. *Training complex cognitive skills: A four-component instructional design model for technical training*. Educational Technology.

Lev Semenovich Vygotsky and Michael Cole. 1978. *Mind in society: Development of higher psychological processes*. Harvard university press.

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*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

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.

James V Wertsch. 1985. *Vygotsky and the social formation of mind*. Harvard university press.

Richard T White. 1973. Research into learning hierarchies. *Review of Educational Research*, 43(3):361–375.

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*.

Caroline E Zsambok and Gary Klein. 2014. *Naturalistic decision making*. Psychology Press.

A Developing Bridge

This section details how we developed the Bridge 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 (the decision options), 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 Examples of Decision Options

This section provides examples for each of decision option. It is split by *error type*, *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 making 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`: The 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`: The 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 Data Processing and Annotation

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

C.1 Data Use

The research team executed Data Use Agreements with both the tutoring provider and school district that outlined the allowable usage of the data to improve instruction in collaboration with an educational agency. Following the FERPA guidelines, we were eligible to engage in secondary data analysis with student data, which is what we did for this study. This study falls under the research team's IRB for conducting research in collaboration with tutoring providers and school district (Protocol #XXXX - redacted due to anonymous submission).

C.2 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.3 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.4 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.

D Prompts

This section contains information on the prompts for gpt-4, gpt-3.5-turbo, and llama-2. We found that we could use similar prompts for gpt-4 and gpt-3.5-turbo, however these prompts had to adapted for llama-2 to mimic its training format[2]. 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). For generating the remediation response, 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. All the prompts include context on the task at the start of the prompt, and the constraints of outputting a JSON-formatted text for the task at the end of the prompt.

D.1 No Decision-Making Condition

Models directly respond, $c_r \sim p(c_r|c_h)$. The prompts for gpt-4 and gpt-3.5-turbo are shown in Figure 4. The prompt for llama-2 is shown in Figure 5 where the formatting is slightly adapted.

D.2 Expert Decision-Making Condition

Models generate with the expert's decisions, $c_r \sim p(c_r|c_h, e, z_{\text{what}}, z_{\text{why}})$. The prompts for gpt-4 and gpt-3.5-turbo are shown in Figure 6. The prompt for llama-2 is shown in Figure 7 where the formatting is slightly adapted. The labels for $e, z_{\text{what}}, z_{\text{why}}$ come from our annotated dataset.

[2] https://gpus.llm-utils.org/llama-2-prompt-template/#notes, https://huggingface.co/blog/llama2#how-to-prompt-llama-2

D.3 Self decision-making condition

LLMs make their own decisions, then generate responses based on them, $c_r \sim p(c_r|c_h, e^{\text{model}}, z^{\text{model}}_{\text{what}}, z^{\text{model}}_{\text{why}})$. Following the decision-making model, we first generate the model's decision on error e^{model} with prompts in Figure 8 (for gpt-4 and gpt-3.5-turbo) and in Figure 9 (for llama-2). Then we generate the model's decision on strategy and intention $z^{\text{model}}_{\text{what}}, z^{\text{model}}_{\text{why}}$ in Figure 10 (for gpt-4 and gpt-3.5-turbo) and in Figure 11 (for llama-2). Finally, we use the previous response generation prompts with decision-making to generate c_r from Section D.2.

D.4 Random Decision-Making Condition

We randomly select a decision for the error, strategy and intention. Then, we use the previous response generation prompts with decision-making to generate c_r from Section D.2.

E Human Evaluations

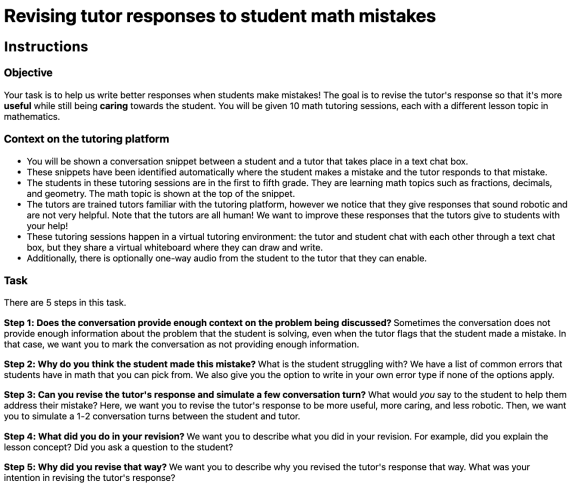
We describe the human evaluation setup, whose results are reported in Section 7.1.

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.

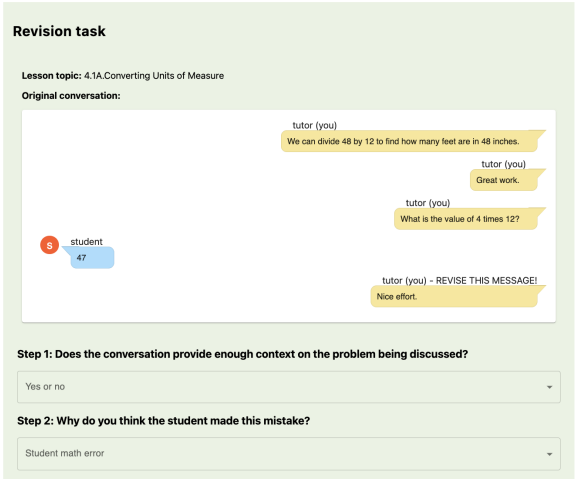
Figure 12 shows an example of the evaluation interface. Specifically, the phrasing for each dimension was:

Which response is more useful?

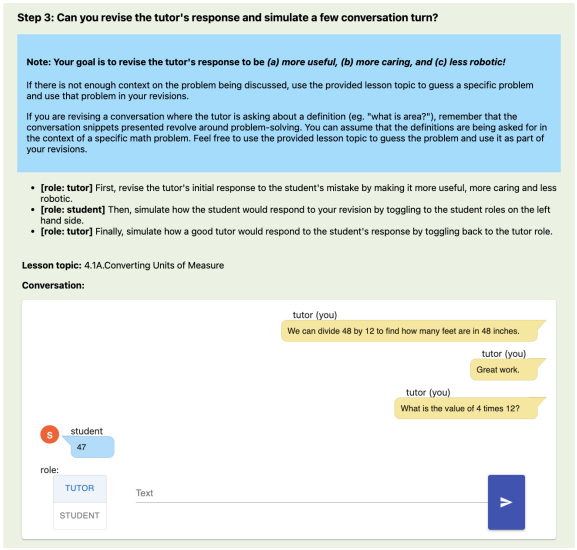
Definition: Useful responses are responses that are productive at advancing the student's understand-



(a) Instructions



(b) Step 1 & 2



(c) Step 3

(d) Step 4 & 5

Figure 3: Annotation interface for collecting decisions and responses.

ing 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.

No Decision-Making Prompt for `gpt-4` and `gpt-3.5-turbo`

```
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}.
{c_h}
tutor (maximum one sentence):
```

Figure 4: **Prompt for the *no decision-making* condition for `gpt-4` and `gpt-3.5-turbo`.** `{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. 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.

No Decision-Making Prompt for `llama-2`

```
### System:
You are an experienced elementary math teacher and you are going to respond to a
student's mistake in a useful and caring way.

### User:
Lesson topic:  {lesson_topic}
Conversation:
{c_h}

### Assistant:
tutor (maximum one sentence):
```

Figure 5: **Prompt for the *no decision-making* condition for `llama-2`.** `{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.

- 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.

F Lexical analysis

Table 4 compares the top-5 bigram usage for `ChatGPT` in all decision-making conditions. Table 5 does the same for `Llama-2-70b-instruct`.

Decision-Making Prompt for `gpt-4` and `gpt-3.5-turbo`

```
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_what} in order to {z_why}.
{c_h}
tutor (maximum one sentence):
```

Figure 6: **Prompt for the *decision-making* condition for `gpt-4` and `gpt-3.5-turbo`.** `{lesson_topic}` is the placeholder for the lesson topic discussed in the conversation. The error, strategy, and intention decisions are included in the prompt where `{e}` is a placeholder for the error type, `{z_what}` for the strategy and `{z_why}` for the intention. Note that each of the decisions are 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.

Decision-Making Prompt for `llama-2`

```
### System:
You are an experienced elementary math teacher and you are going to respond to a
student's mistake in a useful and caring way.

### User:
{e} {z_what} in order to {z_why}.
Lesson topic:  {lesson_topic}
Conversation:
{c_h}

### Assistant:
tutor (maximum one sentence):
```

Figure 7: **Prompt for the *decision-making* condition for `llama-2`.** `{lesson_topic}` is the placeholder for the lesson topic discussed in the conversation. The error, strategy, and intention decisions are included in the prompt where `{e}` is a placeholder for the error type, `{z_what}` for the strategy and `{z_why}` for the intention. Note that each of the decisions are 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.

None + ChatGPT		**Expert + ChatGPT**		**ChatGPT + ChatGPT**		**Random +ChatGPT**	
bigram	log odds	bigram	log odds	bigram	log odds	bigram	log odds
effort_remember	2.30	can_explain	2.01	actually_correct	2.73	thats_close	1.69
effort_double	1.36	great_start	1.88	correct_answer	1.93	example_help	1.58
clarify_mean	1.36	can_tell	1.85	number_baseballs	1.42	can_think	1.47
noticed_mistake	1.36	got_answer	1.63	problem_remember	1.42	can_try	1.47
great_effort	1.34	explain_got	1.53	job_attempting	1.13	good_start	1.36

Table 4: Top 5 bigrams for `ChatGPT`. **`ChatGPT` with *expert decision-making* engages more with the student's problem-solving process, whereas `ChatGPT` with *self decision-making* engages more with the student's answer**.

Determine Error (e) with `gpt-4` and `gpt-3.5-turbo`.

```
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 8: **Prompt to determine error e with `gpt-4` and `gpt-3.5-turbo`.** `{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.

None + Llama		**Expert + Llama**		**Llama + Llama**		**Random +Llama**	
bigram	log odds	bigram	log odds	bigram	log odds	bigram	log odds
user_lesson	6.69	lets_closer	4.02	user_student	5.59	lets_closer	3.52
user_tutor	4.79	closer_look	4.02	student_responds	4.39	closer_look	3.52
teacher_going	3.03	look_problem	2.85	student_understand	3.67	right_track	2.79
experienced_elementary	3.03	problem_break	1.67	response_provide	3.52	youre_right	2.29
going_respond	3.03	groups_objects	1.54	help_student	3.14	student_mistake	1.93

Table 5: Top 5 bigrams for `Llama-2-70b-instruct`.

Determine Error (e) with `llama-2`.

```
### System:
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.

Format your answer as:  [{"answer":  #, "reason":  "write out your reason for picking #
here"}]

### User:
Lesson topic:  {lesson_topic}
Conversation:
{c_h}

### Assistant:
[{"answer":
```

Figure 9: **Prompt to determine error e with `llama-2`.** `{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.

Determine Strategy and Intention ($z_{\text{what}}, z_{\text{why}}$) with `gpt-4` and `gpt-3.5-turbo`.

```
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.

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 10: **Prompt to determine strategy and intention $z_{\text{what}}, z_{\text{why}}$ with `gpt-4` and `gpt-3.5-turbo`.** `{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.

Determine Strategy and Intention ($z_{\text{what}}, z_{\text{why}}$) with `llama-2`.

```
### System:
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.

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)

Format your answer as:  [{"answer":  #, "reason":  "write out your reason for picking #
here"}]

### User:
Lesson topic:  {lesson_topic}
Conversation:
{c_h}

### Assistant:
[{"strategy":
```

Figure 11: **Prompt to determine error $z_{\text{what}}, z_{\text{why}}$ with `llama-2`.** `{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.

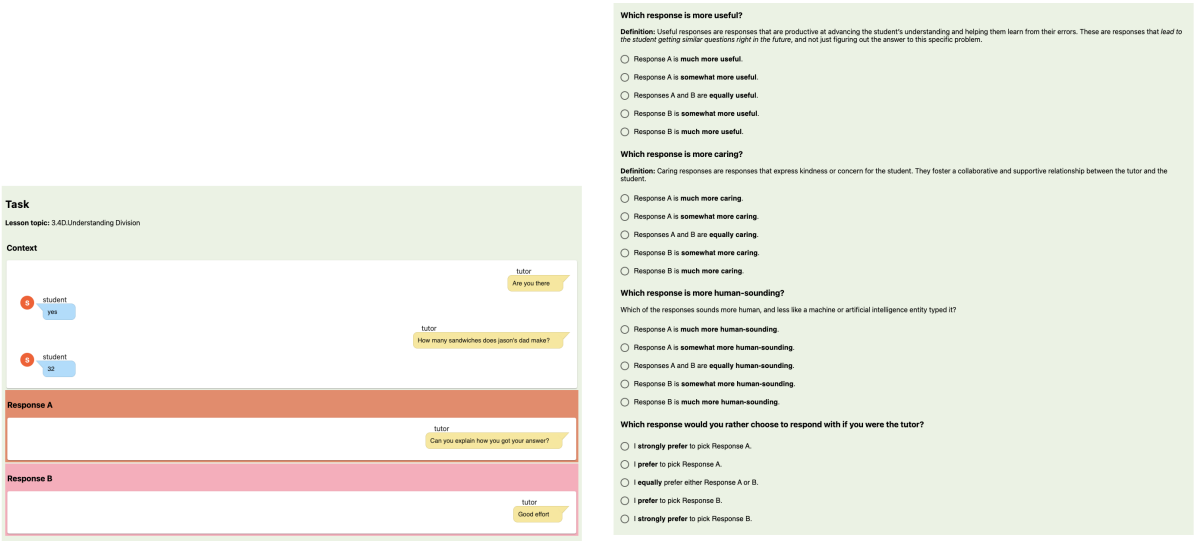


Figure 12: Annotation interface for evaluating Task C remediation responses.