# MathDial: A Dialogue Tutoring Dataset with Rich Pedagogical Properties Grounded in Math Reasoning Problems

**Jakub Macina*** **Nico Daheim*** **Sankalan Pal Chowdhury***
**Tanmay Sinha** **Manu Kapur** **Iryna Gurevych** **Mrinmaya Sachan**

Department of Computer Science, ETH Zürich ETH AI Center
Ubiquitous Knowledge Processing Lab (UKP Lab), Department of Computer Science
and Hessian Center for AI (hessian.AI), TU Darmstadt
Professorship for Learning Sciences and Higher Education, ETH Zürich
jakub.macina@ai.ethz.ch



## Abstract

Although automatic dialogue tutors hold great potential in making education personalized and more accessible, research on such systems has been hampered by a lack of sufficiently large and high-quality datasets. However, collecting such datasets remains challenging, as recording tutoring sessions raises privacy concerns and crowdsourcing leads to insufficient data quality. To address this problem, we propose a framework to semi-synthetically generate such dialogues by pairing real teachers with a large language model (LLM) scaffolded to represent common student errors. In this paper, we describe our ongoing efforts to use this framework to collect MathDial, a dataset of currently ca. 1.5k tutoring dialogues grounded in multi-step math word problems. We show that our dataset exhibits rich pedagogical properties, focusing on guiding students using sense-making questions to let them explore problems. Moreover, we outline that MathDial and its grounding annotations can be used to finetune language models to be more effective tutors (and not just solvers) and highlight remaining challenges that need to be addressed by the research community. We will release our dataset publicly to foster research in this socially important area of NLP.

https://github.com/eth-nlped/mathdial

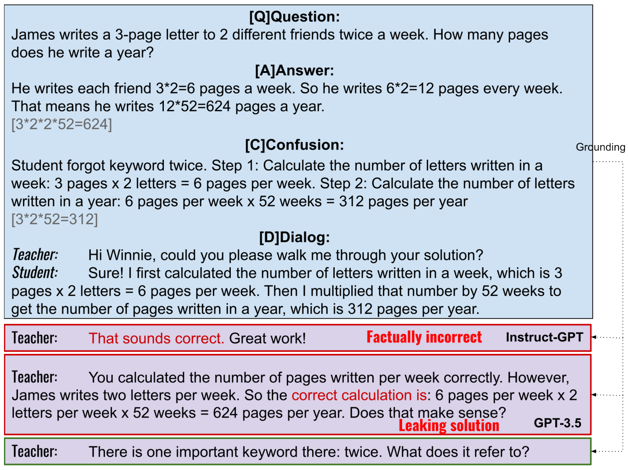


Figure 1: MathDial is grounded in multi-step math reasoning problems, the correct answer, and student confusion. Current models achieve high accuracy in solving such problems but they fail in the task of teaching.

## 1 Introduction

Dialogue tutoring systems have demonstrated significant potential in augmenting learning outcomes across various domains, such as language learning, math, and computer science (Wollny et al., 2021; Ji et al., 2023). However, the progress of scaling such systems is considerably hindered by the lack of high-quality datasets for training effective dialogue models that should provide students with opportunities to learn by scaffolding their learning (Tack and Piech, 2022; Macina et al., 2023). The current datasets are frequently marred with issues like low pedagogical quality, small sizes, or a focus on noisy classroom settings. However, although recording tutoring sessions might seem like a scalable way of collecting data, it bears strong privacy concerns (Demszky and Hill, 2022). On the other hand, crowdsourcing dialogues might provide insufficient quality due to poor annotator training and might furthermore be costly for a sufficiently-sized dataset, since two crowdworkers have to be employed to role-play student and teacher for each dialog.

At the same time, recent advancements in large language models (LLMs) have enabled significant improvements in *generative* dialog systems and

*Equal contribution.

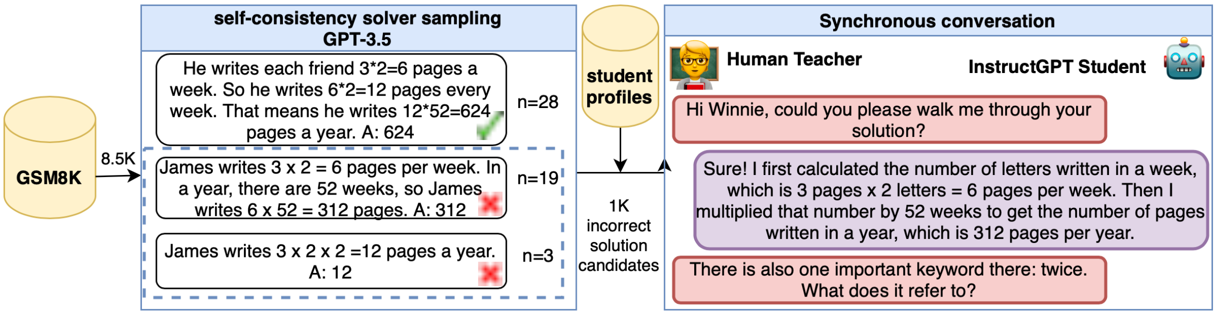


Figure 2: Data collection overview. First, student confusions are oversampled from GPT-3 and sorted by frequency of their occurence. Then, a human teacher synchronously interacts with a student simulated by InstructGPT instructed with a student profile and an incorrect solution.

simultaneously shown great success in reasoning over various educational domains, such as math word problems (Wei et al., 2022; OpenAI, 2023). However, this has not yet translated to improvements in dialogue tutoring systems which lack rich pedagogy and show factually incorrect behaviour for both GPT-3 (Tack and Piech, 2022) and open-source LLMs (Macina et al., 2023). Figure 1 shows examples of generations that tend to not understand student solutions well (Tack and Piech, 2022) and reveal information too early (cf. Section 7.3).

To address these issues, we collect and present a dialogue tutoring dataset, called *MathDial*, that is rich in tutoring quality, which we measure by *equitable tutoring* (Tanner, 2013): the ability to provide opportunities for the student to learn, think and explore potential solutions. For this, we take inspiration from human tutoring strategies (Nye et al., 2014) and active learning approaches in classrooms (Freeman et al., 2014) which show a positive impact on student learning gains. More specifically, we focus on context-sensitive deep reasoning questions which have been shown to be an effective teaching strategy in human tutoring (Nye et al., 2014).

We collect our dataset using a *novel data collection method*, where real teachers are paired with an LLM that simulates student errors deemed representative of real student misunderstandings by the same expert human annotators. MathDial is grounded in math word problems as well as student confusions which provide a challenging testbed for creating faithful and equitable dialogue tutoring models able to reason over complex information. Figure 1 shows an example dialogue, where a teacher scaffolds student learning by asking interactive sense-making questions. Scaffolding (Reiser, 2004) assists students to succeed in tasks that would be otherwise complex for them and let them explore, construct, and self-repair their knowledge which leads to effective human learning (VanLehn, 2011).

Finally, we benchmark different LLMs on our dataset, both using finetuning and prompting. We find that finetuning open-source LLMs on our dataset can make them *significantly more equitable and faithful* compared to just prompting LLMs even with a fraction of parameters. However, there still remains room for improvement, especially for reasoning over more complex problems.

To summarize, our main contributions are:

- A novel framework for semi-synthetic dialogue tutoring dataset creation where real teachers are paired with an LLM that simulates plausible student errors.
- We instantiate our framework in the context of multi-step math word problems to collect an extensive (1.5k dialogues) and high-quality tutoring dataset with rich grounding annotations, called *MathDial*.
- We perform an in-depth analysis of current models outlining the challenges of the dataset and opportunities to use grounding information to build more faithful, verifiable, and equitable tutoring systems.

## 2 Background & Related Work

### 2.1 Dialog Datasets & Collection Methodologies

Research on task-oriented dialogue systems has often focused on customer service conversations. For example, early iterations of the Dialog State Tracking Challenge (DSTC) have dealt with restaurant reservations (Henderson et al., 2014; Gašic et al.,

| Dataset | Domain | Dialogues | Dialogic Pairs | Settings | Grounding Information | Strategies | Bigram Entropy | Uptake | Avg. words per utterance |
|---|---|---|---|---|---|---|---|---|---|
| MathDial (ours) | Math | 1 478 | 7 431 | 1:1 semi-synthetic | confusion, answer | 4 | 3.58 | 0.83 | 17.7 |
| CIMA (Stasaski et al., 2020) | Lang. | 391 | 3 315 | 1:1 role-playing | image, answer | 5 | 3.12 | 0.83 | 13.0 |
| TSCC (Caines et al., 2020) | Language | 102 | 2 013 | 1:1 tutoring | x | 5 | 3.55 | 0.66 | 12.3 |
| TalkMoves (Suresh et al., 2022) | Science | 567 | 9 280 | classroom | x | 10 | 2.93 | 0.67 | 9.6 |
| NCTE (Demszky and Hill, 2022) | Math | 1 660 | 2 348 | classroom | x | 0 | 3.57 | 0.76 | 29.2 |

Table 1: Dialogue tutoring datasets comparison. For NCTE, uptake is calculated on the teacher-student dialogue pairs while bigram entropy is calculated on all teacher utterances. For TalkMoves and TSCC, bigram entropy is calculated on all teacher utterances having more than three words, while uptake is calculated on teacher utterances immediately following student utterances if both have more than three words.

2014). Notably, Wen et al. (2017) collect such dialogues by utilizing the ***Wizard-of-Oz (WoZ)*** paradigm (Kelley, 1984): crowdworkers are connected in a roleplaying game where one of the crowdworkers plays the user interacting with the system, and the other user roleplays as the system itself and might be given additional knowledge sources. Here, the user is generally not aware that they are talking to another crowdworker and not the system. The methodology has also been used to collect the popular MultiWoZ corpus (Budzianowski et al., 2018), which has been revised multiple times (Eric et al., 2020; Zang et al., 2020; Han et al., 2021; Ye et al., 2022) and been extended to document-grounded conversations (Kim et al., 2020). Further corpora collected using the WoZ paradigm include CrossWOZ (Zhu et al., 2020), Taskmaster (Byrne et al., 2019), and open-domain datasets like Wizard-of-Wikipedia (Dinan et al., 2019). Other dialogue datasets have often been collected by having crowdworkers fill dialogue outlines, for example for movie-based (Shah et al., 2018), schema-guided (Rastogi et al., 2020), or high-quality multilingual dialogues (Majewska et al., 2023), or by scraping conversations from the web, for example in DailyDialog (Li et al., 2017), or the Reddit corpus (Dziri et al., 2019).

Different quality issues from crowdsourcing data using non-expert crowdworkers have been reported. For example, Wizard-of-Wikipedia contains many hallucinations in the ground-truth data (Dziri et al., 2022). Similarly, task-oriented conversations suffer from annotation errors and low lexical diversity (Casanueva et al., 2022). More closely related to this work, current tutoring corpora do not provide sufficient tutoring quality (Tack and Piech, 2022; Macina et al., 2023), as we will highlight in the following section. Therefore, in this work we collect a large-scale tutoring dataset that is grounded in math word problems by adapting the WoZ paradigm. Our adaptation differs from related work as it uses a system–in this case a large language model trained for dialogue–to simulate a student in a conversation with real teachers who crowdsource dialogues in collaboration with the model.

### 2.2 Effective Teacher Moves for Dialogue Tutoring & Available Corpora

Former theoretical and empirical work has shown the importance of questioning in human learning (Roscoe and Chi, 2008; Shahriar and Matsuda, 2021; Shridhar et al., 2022) and emphasized conversations to be an essential part of teaching (Freeman et al., 2014). Consequently, prior research has explored which ingredients add value to tutoring conversations. For example, (Nye et al., 2014) show developing expectation-misconception rubrics and asking deep reasoning questions is important, and (Howe et al., 2019) find elaboration and challenging of previous contributions to be positively associated with human learning. This has led to the development of human-authored dialogue tutoring systems. The most prevalent, AutoTutor (Nye et al., 2014), focuses on scaffolding students' natural language explanation to a question or solution to a problem. Recently, several simpler end-to-end rule-based dialogue systems with predefined goals have been proposed (Ruan et al., 2019; Winkler et al., 2020; Cai et al., 2021). However, such systems cannot scale without extensive human authoring and quickly become too complex. Therefore, developing effective automatic tutors *at scale* remains an open problem. While data-driven approaches have seen great success in NLP, to our knowledge only a limited number of tutoring datasets are available: CIMA (Stasaski et al., 2020), TSCC (Caines et al., 2020), TalkMoves (Suresh et al., 2022), and NCTE (Demszky and Hill, 2022). However, all these existing datasets have several limitations such as missing grounding information (TSCC, TalkMoves, NCTE), low tutoring quality (CIMA), small dataset sizes (Macina et al., 2023), or a focus on noisy classroom scenarios.

### 2.3 Synthetic Dialog Data Creation

LLMs have shown increasingly human-like behaviour and therefore naturally found their way as *synthetic dialogue dataset generators* either by being finetuned for inpainting them (Dai et al., 2022) or prompted for synthesis from grounding information (Kim et al., 2022; Chen et al., 2023). The human-like behaviour also manifests in them showing biases in logical reasoning similar to those of humans (Dasgupta et al., 2022; Binz and Schulz, 2023), are on par with gold human annotations for generations tasks (Ziems et al., 2023), and has been used to simulate students for teacher training (Markel et al., 2023). This suggests that one might also be able to rely on LLMs to even create meaningful tutors. However, (Tack and Piech, 2022; Macina et al., 2023) show that they still can not exhibit good performance as *teachers*, as they oftentimes fail to reason correctly about student solutions and reveal answers too quickly, making them good solvers but inadequate tutors.

## 3 MathDial Collection

This section introduces our *novel* end-to-end framework for collecting high-quality tutoring conversations, highlighted in Figure 2. The core idea behind our data collection reverses roles from WoZ simulations: we connect an expert annotator roleplaying a teacher with an LLM that simulates the student [1].

We use this methodology to collect dialogues based on the GSM8k dataset (Cobbe et al., 2021) which provides multi-step math word problems and their solutions with annotated steps. While the problems are simple enough to be understood quickly by teachers, they still contain challenging parts such as solving equations or percentages. For obtaining a tutoring dialogue for a single MWP, the following steps are performed, which we outline in Figure 2.

First, we sample student errors from a large language model in order to provide an initial student confusion that should be resolved in the tutoring dialogue. Therefore, each dialogue is concerned with the solution of exactly one math word problem. Then, we pair a human teacher with the LLM that we ground in a student profile to encourage diverse student behaviour. The profile defines general information about the student, such as their names and errors that frequently occur when solving math word problems, which we base on prior work in algebraic misconceptions (Booth et al., 2017).

We provide the teacher with the math word problem, the incorrect initial student solution that we sample from the LLM, as well as the correct step-by-step solution that is annotated in GSM8k (cf. Figure 6). Then, the teacher uses a sequence of teacher moves in order to clear the student's confusion and guide the student to solving the problem. To promote diverse pedagogical patterns the teacher's moves have to be explicitly selected from a set of four basic moves prior to writing an utterance. However, the teachers themselves can use their expertise to determine a good teaching strategy to tutor the student and are not restricted to specific patterns. We describe these moves further in Section 3.4. The dialogue ends when the teacher assesses that the problem is solved or when a certain time limit is reached. For our collection efforts, we set this limit to five minutes. In addition to the collected dialogues, we obtain a series of metadata that future work can explore for building more effective tutor models. In particular, on top of the math word problem the conversation is grounded in, we provide the step-by-step solution, the misconception of the student as described by the teacher, and annotations on whether the confusion was resolved and typical of 6th grade students.

### 3.1 Teacher Selection

We exclusively recruited professionals with teaching experience through the Prolific platform. Furthermore, to ensure high data quality we filter out all annotators that do not have a 100% completion rate based on more than 500 total submissions. In an initial training phase, annotators read comprehensive guidelines for the task (cf. Section C.1) and then complete a test to assess their understanding of the task on a sample conversation. We only select annotators with 100% test scores. Finally, after manually checking several conversations from each annotator, we select only high-quality participants for the further rounds of data collection, similar to the approach used in (Zhang et al., 2022a). Through this process, we selected 58 expert annotators, of which 47 identified as female and 11 as male. The majority of annotators were nationals of the UK, followed by the USA, Canada, Australia, India, and Germany, with a median age of 40 years

[1] Recall that in WoZ two users are connected, with one simulating a system

| Category | Example Intent | Example Utterance |
|---|---|---|
| **Focus (Scaffolding)** | Seek Strategy | So what should you do next? |
| | Guiding Student Focus | Can you calculate . . . ? |
| | Recall Relevant Information | Can you reread the question and tell me what is . . . ? |
| **Probing (Scaffolding)** | Asking for Explanation | Why do you think you need to add these numbers? |
| | Seeking Self Correction | Are you sure you need to add here? |
| | Perturbing the Question | How would things change if they had . . . items instead? |
| | Seeking World Knowledge | How do you calculate the perimeter of a square? |
| **Telling** | Revealing Strategy | You need to add . . . to . . . to get your answer. |
| | Revealing Answer | No, he had . . . items. |
| **Generic** | Greeting/Fairwell | Hi . . . , how are you doing with the word problem?<br>Good Job! Is there anything else I can help with? |
| | General inquiry | Can you go walk me through your solution? |

Table 2: Non-exhaustive list of intents covered by various utterance categories available to expert annotators. We make use of three broad categories: SCAFFOLDING, TELLING, and GENERIC, of which SCAFFOLDING is divided further into FOCUS and PROBING.

and $10.4\%$ of annotators stemming from non-white ethnicities. Annotators were restricted to having a maximum of five conversations in one annotation session.

### 3.2 Problem & Confusion Selection

We employ an LLM to generate plausible misconceptions on which we base the dialogues. In particular, we pick the most likely incorrect solution generated by ChatGPT (Ouyang et al., 2022) when using chain-of-thought reasoning. To do so, we temperature-sample $N = 50$ reasoning paths for every MWP in GSM8k, with $T = 0.7$ and no top-k truncation, as suggested by (Wang et al., 2023). Then, we group incorrect solutions by their final numeric answer and pick from the group that has the largest cardinality[2]. Indeed, as we will see in the next section, teachers rate the majority of our sampled misconceptions as plausibly made by students. In total, $1181$ incorrect math problem solution candidates are used in the data collection stage.

### 3.3 Generating Student Turns

We use InstructGPT (Ouyang et al., 2022) to generate student turns by prompting the model with the previous dialogue history and additional information that grounds the next turn. In particular, the model is conditioned on the math word problem that needs to be solved, its initial confusion, as well as the student profile (cf. Appendix B for details).

### 3.4 Teacher Moves

For our taxonomy of teacher moves, we follow the least to most specific scaffolding design that is typically used in educational applications because of its empirical effectiveness (VanLehn, 2011; Aleven et al., 2016). Teachers use the least specific–PROBING-to engage students in a deeper understanding of their knowledge, FOCUS to guide the student towards direct progress towards a solution, and the most specific–TELLING-to instruct or reveal parts of the solution e.g. when a student is stuck. We will refer to PROBING and FOCUS as a SCAFFOLDING. Conversational elements with limited pedagogical value are classed as GENERIC. Table 2 gives a list of intents that fall under each of these four categories, along with some examples of each.

### 3.5 Postprocessing

As we are interested in real educational usecases for our tutoring system, we apply a safety filter to filter out conversations with any sensitive content. In particular, we use the Perspective API to filter out conversations containing toxic content (<1%).

## 4 MATHDIAL Analysis

To assess the quality of the dataset, in this section we quantitatively and qualitatively evaluate the tutoring dialogues. As we outline in Table 1, our dataset is significantly larger in terms of both utterance count and the length of teacher utterances than CIMA, which we identify to be the largest one-on-one tutoring dataset that is openly available, thereby filling a gap of sufficiently-sized open-source tutoring corpora (Macina et al., 2023). Furthermore, MATHDIAL exhibits a higher bigram entropy (Zhang et al., 2018) than CIMA, which we use to measure the diversity of teacher utterances. The diversity is on par with transcripts of classroom

[2] cf. Appendix A for details

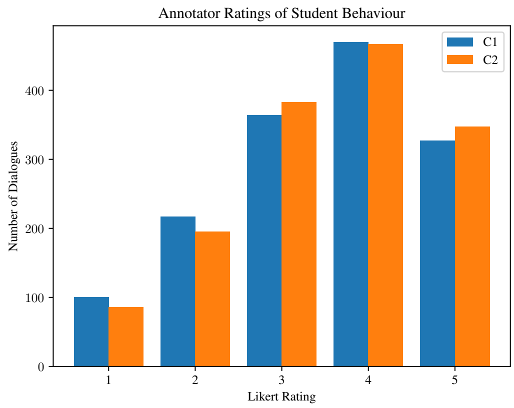


Figure 3: Teacher judgments on the Ability of InstructGPT to simulate students. We find that teachers tend to rate the simulated behaviour as plausible.

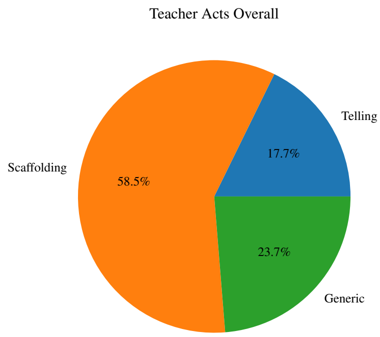


Figure 4: Expert annotators mainly make use of scaffolding moves rather than telling the solution.

and one-to-one tutoring sessions, highlighting the high quality of our data. Moreover, it corroborates the observation that expert annotators tend to create more diverse utterances than untrained crowdworkers (Casanueva et al., 2022). Finally, we measure the Uptake (Demszky et al., 2021) of annotated teacher utterances in our dataset, which indicates how coherent it is with respect to the previous turn. We find that MATHDIAL and CIMA have similar uptake and surpass the other datasets in our comparison.

In the following, we use the annotations provided by the expert crowdworkers to analyze different aspects of MATHDIAL in greater detail: i) how well students can be simulated with LLMs (Section 4.1, ii) which teacher moves teacher employ (Section 4.2) iii) at which stage of the conversation they employ them (Section 4.3), and finally iv) how often student confusions are resolved (Section 4.4)

### 4.1 How well can LLMs simulate students?

Since our collection methodology relies on using LLMs to simulate students, it is crucial that the model can do so realistically to create high-quality data. To assess the quality, we ask teachers to judge the following criteria:

C1: The student had a confusion like a typical 6th grade student.

C2: The student interactions were reasonable and expected of 6th grade students.

The annotators were asked to rate both on a 5-point Likert scale, ranging from 1–*very unlikely*–to 5–*very likely*. Figure 3 shows that annotators rate the majority of simulations by the model positively, confirming that it is possible to use LLMs to simulate students that teachers also conceive to be plausible and representative of 6th grade students. We release these annotations as a part of our final dataset.

Moreover, we asked annotators to distinguish conceptual errors from simple calculation errors. While arithmetic errors may be resolved through the usage of calculators, conceptual errors are more likely to require scaffolding student understanding by tutors. Annotators identified 75% of the confusions as *conceptual*, and accordingly a quarter of the data as basing on arithmetic mistakes.

### 4.2 Which teaching moves do annotators choose?

We have utterance level annotations from the teachers for the utterance categories described in Table 2. Figure 4 shows the overall distribution of teacher moves. We can see that SCAFFOLDING is used most frequently in almost 60% of utterances, followed by GENERIC, while TELLING is the rarest. To validate the quality of these annotations, we sampled 17 conversations consisting of 102 teacher utterances and ask two expert annotators to annotate their moves independently. While we obtain a Fleiss' $\kappa$ of 0.669 for combining PROBING and FOCUS the agreement between expert annotators and teacher annotations is low if they are kept separately. Nevertheless, this is in line with similar works (Kelly et al., 2020) that have shown low inter-annotator agreement for moves in classroom settings.

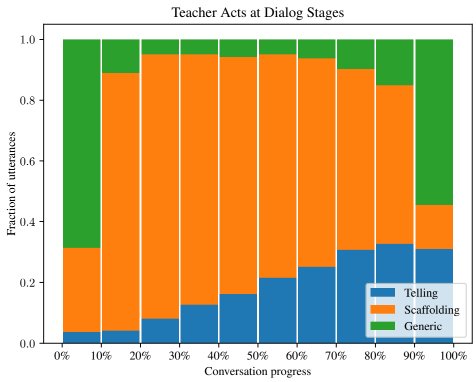


Figure 5: Relative proportions of various teacher acts at different stages of the dialog. Teachers tend to start with SCAFFOLDING but tend to use TELLING more as the conversation progresses.

### 4.3 When do teachers choose which move?

In this section, we analyze at which stages teachers choose particular moves towards the resolution of the student confusion. Figure 5 shows how often annotators use various acts as conversations progress, measured by how frequently they occur at certain *relative* stages of the conversation. We find that the initial utterance by the teacher tends to be generic, mainly to provide an opener or ask the student to repeat the solution attempt. During the conversation, teachers mainly resort to scaffolding utterances that either probe the student or contain focus questions. As the conversation progresses, teachers then resort to telling more often, as students might get stuck at some specific subproblem and might not be able to resolve it on their own. Furthermore, this helps to keep them engaged in the conversation (VanLehn, 2011).

### 4.4 How often can student confusion be resolved?

Finally, we would like to know how often the teachers are able to resolve the confusions of the student in our collected dialogues. The expert annotators assessed that in $86.69\%$ cases they were able to do so: in $72.19\%$ of the conversations by using just sensemaking questions, and only in $14.50\%$ of cases by revealing the majority of the answer. The small percentage of unresolved confusions seems to predominantly be caused by InstructGPT's inability to deal with larger contexts. However, we deem them still useful, as they can be used to train downstream models on when to seek human intervention, for example.

## 5 Modeling Tutors With MATHDIAL

We focus our intial studies on MATHDIAL on the task of tutor response generation. The goal of tutor response generation is to model the teacher in a dialogue by generating appropriate follow-up turns that guide the student towards solving the problem. That is, given a tutoring dialogue history $u_1^T$, grounding knowledge $\mathcal{K}$, and a teacher move $\mathcal{A}$, we wish to generate a continuation of the dialogue $u_{T+1} \subset \mathcal{V}^*$. Therefore, we study locally-normalized neural language generators of the form

$$p_{\boldsymbol{\theta}}(u_{T+1} \mid u_1^T, \mathcal{K}, \mathcal{A}) = \prod_{n=1}^{N_{T+1}} p_{\boldsymbol{\theta}}([u_{T+1}]_n \mid [u_{T+1}]_1^{n-1}, u_1^T, \mathcal{K}, \mathcal{A}), \quad (1)$$

that are parametrized by a set of parameters $\boldsymbol{\theta}$.

In the following, we evaluate different parametrizations based on pretrained Transformer models (Vaswani et al., 2017a). Furthermore, we study the influence of different kinds of grounding information $\mathcal{K}$ that are annotated per dialogue in DialogMath. Specifically, we can ground the model in i) the math word problem in question, ii) the ground-truth solution with step-wise annotations, iii) and the initial student misconception.

## 6 Experiments

For tutor response generation, we experiment with multiple state-of-the-art pretrained Transformer (Vaswani et al., 2017b) models. Specifically, we use BART (Lewis et al., 2020), an encoder-decoder model trained on denoising tasks, flan-t5 (Chung et al., 2022), a text-to-text trained encoder-decoder model based on T5 (Raffel et al., 2020) that was further finetuned on the instruction-following flan collection. Furthermore, we make use of OPT (Zhang et al., 2022b) a decoder-only model trained on a next-token-prediction task. Finally, we compare these models against GPT-3 (Brown et al., 2020), which we do not finetune but rather prompt in a zero-shot manner. Furthermore, we use flanT5, BART, and GPT-3 to generate rationales that describe the student's mistakes. We train all models for 10 epochs with an initial learning rate of $6.25e-5$, linear learning rate decay without warmup and by making use of the checkpoints provided by the transformers library (Wolf et al., 2020).

| Model | MathDial | | | | | MathDial$_{\text{seen}}$ | | MathDial$_{\text{unseen}}$ | |
|---|---|---|---|---|---|---|---|---|---|
| | sBLEU | BERT F1 | KF1 | BERT F1 | Uptake | sBLEU | KF1 | sBLEU | KF1 |
| | $(u_{T+1}, \hat{u}_{T+1})$ | | $(u_{T+1}, \mathcal{K})$ | | $(u_T, u_{T+1})$ | | | | |
| finetuned | | | | | | | | | |
| BART$_{139M}$ | 6.0 | 54.0 | 19.9 | 50.3 | 0.89 | 6.7 | 19.5 | 5.0 | 20.5 |
| BART$_{406M}$ | 5.3 | 50.7 | 19.7 | 50.0 | 0.88 | 5.8 | 19.7 | 4.6 | 19.6 |
| T5$_{250M}$ | 5.9 | 50.6 | **31.3** | **56.6** | **0.95** | 6.4 | **30.6** | 5.0 | **32.4** |
| T5$_{780M}$ | 6.2 | 51.7 | 28.7 | 55.6 | 0.94 | 6.7 | 28.7 | 5.5 | 28.1 |
| flanT5$_{250M}$ | 6.7 | 52.8 | 28.2 | 54.9 | 0.92 | 7.0 | 29.5 | 5.4 | 28.8 |
| flanT5$_{780M}$ | **8.1** | **55.1** | 24.6 | 53.1 | 0.90 | **8.9** | 23.9 | **6.8** | 25.6 |
| flanT5$_{3B}$ | 7.3 | 53.9 | 22.4 | 50.4 | 0.90 | 7.6 | 22.0 | 6.7 | 22.7 |
| OPT$_{125M}$ | 3.8 | 51.9 | 13.1 | 44.8 | 0.82 | 4.5 | 13.2 | 2.7 | 13.1 |
| OPT$_{1.3B}$ | 4.2 | 52.0 | 12.3 | 44.4 | 0.79 | 4.3 | 12.2 | 2.9 | 9.6 |
| 0-shot | | | | | | | | | |
| GPT-3 | 2.5 | 49.0 | 22.8 | 50.3 | 0.92 | 2.8 | 22.8 | 2.1 | 22.7 |

Table 3: Results of finetuned and zero-shot prompted models on the tutor response generation task. We find that i) models finetuned on our dataset can outperform much larger prompted models, ii) there is still a gap in terms of generalization, iii) simply scaling the same pretrained model does not immediately improve results.

We split our data along conversations such that 80% of the conversations are contained in the training data and 20% in the test set. Our test set contains around 2/3 seen problems, where the model has seen at least one conversation dealing with it during training, and 1/3 of unseen problems, which the model has not encountered in training. We use these two subsets to benchmark the generalization abilities of our models.

We assess our models using the sacrebleu (Post, 2018) implementation of BLEU (Papineni et al., 2002), as well as BERTScore (Zhang* et al., 2020) [3]. Furthermore, in line with previous works (Dziri et al., 2022; Daheim et al., 2023) we report BERTScore and token-level F1 between generated utterance and math word problem as a proxy for faithfulness. However, we note that an increase in overlap may also indicate more telling, which can be undesirable. Finding good evaluation metrics for assessing the faithfulness of dialogue tutors hence remains an open problem.

# 7 Results

## 7.1 Tutor Response Generation

Table 3 shows the results we obtain with different models pretrained on MathDial. In general, the automatic metrics are significantly lower than results reported with state-of-the-art models on other datasets, which we hypothesize to have two causes. First of all, tutoring is an inherently ambiguous task that admits multiple feasible strategies to teach the same concept and furthermore, teachers may express the same intent in multiple different ways. Secondly, the concepts in which the models are grounded are significantly more challenging in MathDial, as the model has to actively reason over math word problems and different kinds of information. On the other hand, the grounding information given in other grounded dialog datasets oftentimes already determines the response to a significant extent, which then is only a rewritten form of it.

Scaling models in terms of their parameter size is not directly reflected in improved metrics, indicating that just training larger models on more data might not suffice to build tutors on MathDial. Nevertheless, as shown in BERT F1 between response and grounding information, as well as their lexical overlap, smaller models appear to rely more on the grounding information and might be less able to paraphrase. We can note that instruction tuning of models (flanT5) improves their performance also on tutoring data and that encoder-decoder models appear to perform better than decoder-only models. Nevertheless, the comparison does not control for model and pretraining data size. In terms of BLEU score all models perform better on the seen data and some models appear to rely more on the grounding information on the unseen data. Uptake is generally high and for different models even higher than the ground-truth annotations. Finally, finetuned models tend to outperform zero-shot prompted GPT despite significantly smaller lower parameter counts.

## 7.2 Influence of grounding information

Table 4 shows results obtained with flanT5$_{250M}$ with varying grounding information, compared to an ungrounded model. We can see that, while the additional metadata increases faithfulness metrics,

[3] We use the deberta-large-mnli checkpoint

| | sBLEU | BERT F1 | F1 | BERT F1 |
|---|---|---|---|---|
| | $(u_{T+1}, \hat{u}_{T+1})$ | | $(u_{T+1}, \mathcal{K})$ | |
| flanT5$_{250M}$ | 6.9 | **54.0** | 23.3 | 52.3 |
| + question | 6.7 | 53.8 | **27.3** | **54.9** |
| + incorrect solution | 6.7 | 53.5 | 23.9 | 52.7 |
| + ground-truth | **7.2** | 53.8 | 24.6 | 52.8 |

Table 4: Ablation on the influence of grounding information used by the model.

other metrics remain in a similar range. Of the different information, the model seems to make use of the question the most, followed by the step-by-step solution and the incorrect solution attempt.

### 7.3 Human Evaluation

| Model | Coherence | Correctness | Equitable |
|---|---|---|---|
| flanT5$_{780M}$ | 2.48 | 0.89 | **2.34** |
| flanT5$_{3B}$ | 2.61 | **0.90** | 2.33 |
| OPT$_{1.3B}$ | 2.27 | 0.73 | 2.14 |
| GPT3 | **2.84** | 0.54 | 1.53 |
| Ground-truth | 2.84 | 0.95 | 2.57 |

Table 5: Our human evaluation shows that finetuning models on MATHDIAL increases their performance in terms of correctness and equitable tutoring.

Finally, we assess our models using a human evaluation according to three criteria: i) *Coherence*: how coherent the teacher response is with respect to the preceding dialogue, ii) *Correctness*: whether it is in itself correct, and iii) *Equitable* tutoring. We use three expert annotators that judge $n = 50$ responses each. First of all, we find that the ground-truth annotations are of high-quality with high ratings in terms of all criteria. Then, notably, we find that small finetuned models perform much better in terms of correctness and equitable tutoring than a prompted large language model (GPT3) with a significantly larger parameter count, indicating that MATHDIAL indeed allows for training more effective tutors. The automatic metrics are only partly confirmed, as flanT5$_{3B}$ appears to be more coherent than flanT5$_{780M}$ despite receiving lower BLEU scores. Nevertheless, the lower performance of OPT$_{1.3B}$ is confirmed.

## 8 Discussion & Conclusion

In this paper, we have introduced a novel methodology for dialogue dataset collection, where LLMs are used to realistically stimulate students and their mistakes and are paired with real teachers to scalably collect conversations on problem-solving. By instantiating this method on GSM8k (Cobbe et al., 2021), we have collected MATHDIAL, a novel dataset of 1.5k tutoring conversations grounded in math word problems. The dataset fills a crucial gap towards building effective dialogue tutors at scale, providing a significantly larger amount of dialogues than other available corpora in one-to-one tutoring. Our methodology leads to a pedagogically rich dataset that aligns with equitable tutoring practices and research on how to scaffold student understanding. We think the dataset not only progresses the field of dialogue tutoring but dialogue systems as a whole, as it provides a tough testbed for more faithful dialogue models, since modeling MATHDIAL effectively requires a significant amount of reasoning over nontrivial mathematical concepts. Moreover, the rich grounding annotations enable the building more faithful and equitable dialogue tutoring systems (and not just solvers).

We benchmark currently available open-source language models on the task of tutor response generation. We show that smaller models finetuned on our MathDial can significantly surpass the performance of much larger prompted LLMs. Nevertheless, different challenges remain. Namely, dialogue models still require better reasoning over student solutions and there is still room for improvement in modeling teaching strategies. Furthermore, all models struggle to correctly describe the misunderstanding of the student.

## 9 Limitations

In this work, we used ChatGPT to simulate student confusions. However, we acknowledge that these models have a limited understanding of human learning and this is a key limitation in our dataset – certain kinds of student confusions may be under- or over-represented in our dataset. Future work can focus on addressing this limitation. Furthermore, in our setup, teachers were interacting with ChatGPT role-playing as a student. However, it is possible that some teachers might have learned to interact with the student model in a different way than they would do in the classroom. Furthermore, it is also possible that some teachers may have lost motivation when found out they are not interacting with real students, leading to lower data quality.

The methodology to collect the dataset was instantiated just for the domain of math reasoning. The collection of additional domain-specific datasets is necessary to further generalize the effectiveness of the methodology.

## References


Vincent Aleven, Elizabeth A McLaughlin, R Amos Glenn, and Kenneth R Koedinger. 2016. Instruction based on adaptive learning technologies. *Handbook of research on learning and instruction*, 2:522–560.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Julie L Booth, Kelly M McGinn, Christina Barbieri, and Laura K Young. 2017. Misconceptions and learning algebra. *And the rest is just algebra*, pages 63–78.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4516–4525, Hong Kong, China. Association for Computational Linguistics.

William Cai, Josh Grossman, Zhiyuan Jerry Lin, Hao Sheng, Johnny Tian-Zheng Wei, Joseph Jay Williams, and Sharad Goel. 2021. Bandit algorithms to personalize educational chatbots. *Machine Learning*, 110(9):2389–2418.

Andrew Caines, Helen Yannakoudakis, Helena Edmondson, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2020. The teacher-student chatroom corpus. In *Proceedings of the 9th Workshop on NLP for Computer Assisted Language Learning*, pages 10–20.

Inigo Casanueva, Ivan Vulić, Georgios Spithourakis, and Paweł Budzianowski. 2022. NLU++: A multi-label, slot-rich, generalisable dataset for natural language understanding in task-oriented dialogue. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1998–2013, Seattle, United States. Association for Computational Linguistics.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. PLACES: Prompting language models for social conversation synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Nico Daheim, Nouha Dziri, Mrinmaya Sachan, Iryna Gurevych, and Edoardo M. Ponti. 2023. Elastic weight removal for faithful and abstractive dialogue generation.

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. In *International Conference on Machine Learning*, pages 4558–4586. PMLR.

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*.

Dorottya Demszky and Heather Hill. 2022. The ncte transcripts: A dataset of elementary math classroom transcripts. *arXiv preprint arXiv:2211.11772*.

Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori Hashimoto. 2021. Measuring conversational uptake: A case study on student-teacher interactions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Inter-*

*national Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1638–1653, Online. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Augmenting neural response generation with context-aware topical attention. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 18–31, Florence, Italy. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M Ponti, and Siva Reddy. 2022. Faithdial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Scott Freeman, Sarah L Eddy, Miles McDonough, Michelle K Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. 2014. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the national academy of sciences*, 111(23):8410–8415.

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

Milica Gašic, Dongho Kim, Pirros Tsiakoulis, Catherine Breslin, Matthew Henderson, Martin Szummer, Blaise Thomson, and Steve Young. 2014. Incremental on-line adaptation of pomdp-based dialogue managers to extended domains. *In Proceedings on InterSpeech*.

Ting Han, Ximing Liu, Ryuichi Takanabu, Yixin Lian, Chongxuan Huang, Dazhen Wan, Wei Peng, and Minlie Huang. 2021. Multiwoz 2.3: A multi-domain task-oriented dialogue dataset enhanced with annotation corrections and co-reference annotation. In *Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part II 10*, pages 206–218. Springer.

Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Christine Howe, Sara Hennessy, Neil Mercer, Maria Vrikki, and Lisa Wheatley. 2019. Teacher–student dialogue during classroom teaching: Does it really impact on student outcomes? *Journal of the learning sciences*, 28(4-5):462–512.

Hyangeun Ji, Insook Han, and Yujung Ko. 2023. A systematic review of conversational ai in language education: focusing on the collaboration with human teachers. *Journal of Research on Technology in Education*, 55(1):48–63.

John F Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, 2(1):26–41.

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.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, et al. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv preprint arXiv:2212.10465*.

Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tur. 2020. Beyond domain APIs: Task-oriented conversational modeling with unstructured knowledge access. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 278–289, 1st virtual meeting. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Margarita Limón. 2001. On the cognitive conflict as an instructional strategy for conceptual change: A critical appraisal. *Learning and instruction*, 11(4-5):357–380.

Jakub Macina, Nico Daheim, Lingzhi Wang, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Opportunities and challenges in neural dialog tutoring. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2357–2372, Dubrovnik, Croatia. Association for Computational Linguistics.

Olga Majewska, Evgeniia Razumovskaia, Edoardo M Ponti, Ivan Vulić, and Anna Korhonen. 2023. Cross-lingual dialogue dataset creation via outline-based generation. *Transactions of the Association for Computational Linguistics*, 11:139–156.

Julia M Markel, Steven G Opferman, James A Landay, and Chris Piech. 2023. Gpteach: Interactive ta training with gpt based students.

Benjamin D Nye, Arthur C Graesser, and Xiangen Hu. 2014. Autotutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, 24:427–469.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Brian J. Reiser. 2004. Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *Journal of the Learning Sciences*, 13(3):273–304.

Rod D Roscoe and Michelene TH Chi. 2008. Tutor learning: The role of explaining and responding to questions. *Instructional science*, 36:321–350.

Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L Murnane, Emma Brunskill, and James A Landay. 2019. Quizbot: A dialogue-based adaptive learning system for factual knowledge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. 2018. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*.

Tasmia Shahriar and Noboru Matsuda. 2021. “can you clarify what you said?”: Studying the impact of tutee agents’ follow-up questions on tutors’ learning. In *Artificial Intelligence in Education: 22nd International Conference, AIED 2021, Utrecht, The Netherlands, June 14–18, 2021, Proceedings, Part I 22*, pages 395–407. Springer.

Kumar Shridhar, Jakub Macina, Mennatallah El-Assady, Tanmay Sinha, Manu Kapur, and Mrinmaya Sachan. 2022. Automatic generation of socratic subquestions for teaching math word problems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4136–4149, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Katherine Stasaski, Kimberly Kao, and Marti A Hearst. 2020. Cima: A large open access dialogue dataset for tutoring. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–64.

Abhijit Suresh, Jennifer Jacobs, Margaret Perkoff, James H Martin, and Tamara Sumner. 2022. Fine-tuning transformers with additional context to classify discursive moves in mathematics classrooms. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 71–81.

Anaïs Tack and Chris Piech. 2022. The AI teacher test: Measuring the pedagogical ability of blender and GPT-3 in educational dialogues. In *Proceedings of the 15th International Conference on Educational Data Mining*, pages 522–529, Durham, United Kingdom. International Educational Data Mining Society.

Kimberly D Tanner. 2013. Structure matters: twenty-one teaching strategies to promote student engagement and cultivate classroom equity. *CBE—Life Sciences Education*, 12(3):322–331.

Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational psychologist*, 46(4):197–221.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *ICLR 2023*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 2020. Sara, the lecturer: Improving learning in online education with a scaffolding-based conversational agent. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Sebastian Wollny, Jan Schneider, Daniele Di Mitri, Joshua Weidlich, Marc Rittberger, and Hendrik Drachsler. 2021. Are we there yet?-a systematic literature review on chatbots in education. *Frontiers in artificial intelligence*, 4:654924.

Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2022. MultiWOZ 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 351–360, Edinburgh, UK. Association for Computational Linguistics.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. Association for Computational Linguistics.

Lining Zhang, João Sedoc, Simon Mille, Yufang Hou, Sebastian Gehrmann, Daniel Deutsch, Elizabeth Clark, Yixin Liu, Miruna Clinciu, Saad Mahamood, and Khyathi Chandu. 2022a. Needle in a haystack: An analysis of finding qualified workers on mturk for summarization.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022b. Opt: Open pre-trained transformer language models.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. *Advances in Neural Information Processing Systems*, 31.

Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. 2020. Crosswoz: A large-scale chinese cross-domain task-oriented dialogue dataset. *Transactions of the Association for Computational Linguistics*, 8:281–295.

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 Sampling student incorrect solution candidates

We follow the GSM8k reasoning format and prompt InstructGPT with a 2-shot prompt. Given a prompt and a math word problem, we sample $n$ reasoning paths $r_i$ solutions from the model. We parse the first numerical answer $a_i$ after the model generated "####" which represents the final result. Most of the generated outputs have this format and we discard all generations not following it. We sample $N = 50$ reasoning path candidates using the same settings as suggested by (Wang et al., 2023). After sampling multiple reasoning pairs and corresponding answer pairs $(r_i, a_i)$ we use a majority vote over $a_i$ which does not lead to a ground truth answer $a$: $\arg\max_a \sum_{i=1}^{n} \mathbb{1}(a_i \neq a)$. We select problems with at most four solution steps. Since our initial experiments show the occurence of rounding errors, which related work finds to be more common in LLMs than humans (Frieder et al., 2023), we limit them by discarding confusions that are within $0.1$ of the original solution. Moreover, to filter out other simple calculation errors which are not interesting from a learning standpoint we parse all the intermediate equations which are in the format $<< a \times b = c >>$ and use a calculator to check for inconsistencies.

The full prompt used is:

```
Q: Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?
A: Natalia sold 48/2 = «48/2=24»24 clips in May. Natalia sold 48+24 = «48+24=72»72 clips altogether in April and May. #### 72
Q: Weng earns $12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn?
A: Weng earns 12/60 = «12/60=0.2»0.2 per minute. Working 50 minutes, she earned 0.2 x 50 = «0.2*50=10»10. #### 10
```

Of the problems in the GSM8k Train set, $5684$ problems were queried after eliminating problems with more than 5 steps in the solution. This cost a total of $27,425,152$ ChatGPT Tokens and yielded $2,313$ problems with at least one wrong solution. We then eliminated student solutions having fewer than $300$ characters (having too few characters makes it harder to pinpoint where exactly the error occurred) or more than $500$ characters (longer solutions require annotators to spend more time understanding the error), leaving us with $1,379$ wrong solutions. Finally, we eliminate problems where all $50$ or $49$ out of $50$ proposed solutions have the same(wrong) final answer, leaving us with our final set of $1,161$ problems.

# B Details of model for generating student responses

## B.1 Prompt

We use InstructGPT with the following prompt using temperature-sampling with $T = 0.4$ and no top-k truncation:

```
Student Persona: (STUDENT PERSONA)\n\n
Math problem: (MATH PROBLEM)\n\n
Student solution: (STUDENT SOLUTION)\n\n
Context: (STUDENT NAME) thinks their answer is correct. Only when the teacher provides several good reasoning questions, (STUDENT NAME) understands the problem and corrects the solution. (STUDENT NAME) can use a calculator and thus makes no calculation errors. Send EOM tag at end of the student message.\n\n
(DIALOG HISTORY)
```

## B.2 Student characteristics

List of all student characteristics based on prior work studying misconceptions in learning algebra (Booth et al., 2017):

- has a problem with understanding what steps or procedures are required to solve a problem.
- has a problem with understanding underlying ideas and principles and a recognition of when to apply them.
- struggle most with understanding what the problem is asking them to do.
- has difficulty determining which pieces of information are relevant and which are irrelevant to solving the problem.
- struggle to put the numbers in the correct order in the equation or determine the correct operation to use.
- struggle to recognize the problem type and therefore do not know what strategy to use to solve it.

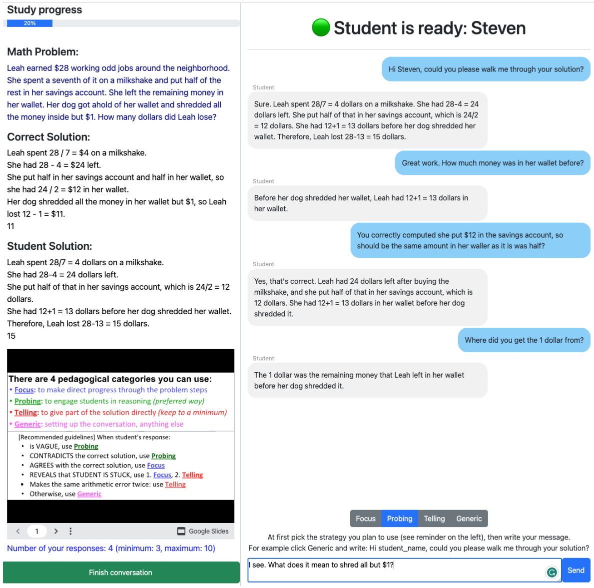


Figure 6: Web interface for collecting dialogue tutoring conversations. The left panel shows math word problem, correct solution, and student solution. The right panel contains conversation history, a panel for selecting the category of response, and a text area to send a response to the student. After clicking Send, the student model is immediately invoked using an internal API call.

### B.3 Common error cases

We manually screened conversations and teacher feedback. The most common problem was the occurence of simple arithmetic errors (e.g. 7-2=6) and inconsistent student behaviour (e.g. student returning back to the incorrect answer after figuring out the correct one in the previous utterance). These errors were captured in the teacher quality evaluation of student behaviour.

## C Data collection details

We use Prolific for data collection and hire annotators with teaching experience. To ensure the data quality we filter only annotators with 100% completion rate with more than 500 total submissions. All the payments to the annotators exceeded the US federal minimum wage and the final batch of annotators were paid the equivalent of $10/hour. The data collection interface is shown in Figure 6.

**Training phase** We let annotators read best practices on how to have a productive conversation with students (cf. Section C.1) and tested them on their understanding of our task afterwards. We started the data annotation with all the annotators able to successfully pass the test. Later, we manually checked several conversations by each annotator in terms of quality and used diverse scaffolding questions. We selected only high-quality participants using these criteria for the further rounds of data collection.

### C.1 Annotation Guidelines

Teachers were instructed to have a one-on-one tutoring session with different 6th-grade students. They were told that students received a math word problem for homework and submitted their solutions beforehand. In a tutoring conversation, teachers were asked to go through the student's solution and try to let the student understand using a series of sense-making questions to support student reasoning and learning. Specifically, they were instructed to not just correct student solutions by telling what's correct/incorrect, but to give students the opportunity to explore the problem with a focus on core aspects it, such as their chosen strategy. However, as the goal is to focus on conceptual errors, they were allowed to let students use calculators or correct their arithmetic mistakes.

### C.2 Schema for Teacher Moves

Scaffolding (Reiser, 2004) assists students to succeed in tasks that would otherwise be complex and differentiates between guidance (e.g. decomposing problem, clarifying) from cognitive activation (e.g. causing cognitive conflicts, activating prior knowledge(Limón, 2001)). The effective teacher moves to scaffold students' understanding have been studied extensively by analyzing and annotating real human tutoring conversations (Nye et al., 2014; VanLehn, 2011). Experienced teachers can through natural language guide students' focus and uncover misconceptions (Nye et al., 2014). The teacher moves in the form of scaffolding support student understanding by asking open-ended questions, activating their prior knowledge, or causing cognitive conflicts (Limón, 2001). A teacher asking sense-making questions provides learning opportunities for students to actively construct their knowledge. However, at the same time asking only difficult questions could lead to a loss of learner motivation and potentially the end of the dialogue. On the other hand, only constantly revealing answers does not lead to long-term learning.

Table 2 refers to the details of teacher moves used during annotation and their underlying learning mechanism. Most importantly, scaffolding questions that are productive for long-term learning are *Focus* and *Probing*. On the other hand, *Telling*

```
Prompt:
A proficient educator assists a learner by providing just a few clues to help him/her in the correct direction. The
student should take the lead in reasoning and computations. Generate a tutoring conversation between a math teacher and
an 8th-grade student Leslie unable to solve the following math word problem correctly.

Math Problem: Jeremy decides to take Tylenol because of the pain he is experiencing. He decides to take 1000 mg every 6
hours for 2 weeks.  How many pills do they take if they are each 500 mg?

Correct solution: He takes 1000/500=<<1000/500=2>>2 pills each time. He takes them 24/6=<<24/6=4>>4 times a day. So he
takes 4*2=<<4*2=8>>8 pills per day. He takes pills for 2*7=<<2*7=14>>14 days. So he takes 8*14=<<8*14=112>>112 pills

InstructGPT:
Teacher: Hi Leslie, it looks like you're having some trouble with this math word problem. Can you tell me what you're
struggling with?

Leslie: Yeah, I'm not sure how to figure out how many pills Jeremy needs to take.

Teacher: Alright, let's break down the problem and see if we can solve it step by step. First, let's think about how much
Tylenol Jeremy needs to take each day. He's taking 1000 mg every 6 hours, so how many times does he need to take it each
day?

Leslie: He needs to take it 3 times a day because 24 / 6 = 3.

Teacher: That's correct! Now let's figure out how much Tylenol he takes each day in total. If he takes 1000 mg each time,
and he takes it 3 times a day, how much does he take in total each day?

Leslie: He takes 3000 mg each day because 1000 x 3 = 3000.

Teacher: Excellent! Now let's figure out how many pills he needs to take each day. Each pill is 500 mg, so how many pills
does he need to take each day to get his 3000 mg?

Leslie: He needs to take 6 pills each day because 3000 / 500 = 6.

Teacher: That's correct! Now let's think about how many pills he needs to take in total over the 2 weeks. If he needs to
take 6 pills each day, and there are 7 days in a week, how many pills does he need to take in total over 2 weeks?

Leslie: He needs to take 84 pills in total because 6 x 7 x 2 = 84.

Teacher: Great job, Leslie! You've done a fantastic job breaking down the problem and solving it step by step. Well done!
```

Figure 7: In our initial pilot study we observed that synthetic data generation has strictly followed the same structure of only asking next-step questions (highlighted in yellow) and was prone to inconsistencies in factual correctness and order of steps (highlighted in red).

represents giving out the partial or full answer to the student and should be mostly used when a student is stuck.

In summary, *Focus* comprises of all conversation elements that direct the student towards the solution without actually giving out any of the solution, while *Probing* attempts to develop reasoning skills and world knowledge relevant to the problem, but not necessarily specific to the given problem. *Telling* is giving out parts of the solution, either calculations or strategy or both. All other conversational elements, including trying to understand what the student has already tried, fall under *Generic*.

# D Initial pilots

Example conversation of synthetic data generation is shown in Figure 7.

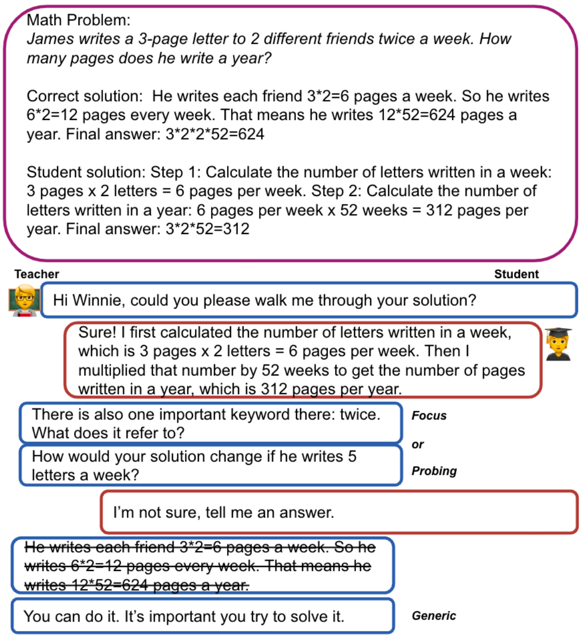


Figure 8: Example of 1:1 dialog tutoring grounded in the math word problem. Each conversation is grounded in the correct solution and student solution with corresponding reasoning graph enabling verifying student solution and tracking the state of the problem-solving.

| Context | Conversation |
| --- | --- |
| *Q*: Sam memorized six more digits of pi than Carlos memorized. Mina memorized six times as many digits of pi as Carlos memorized. If Mina memorized 24 digits of pi, how many digits did Sam memorize?<br><br>*Ground truth*: Carlos memorized 24/6=4 digits of pi. Sam memorized 4+6=10 digits of pi. Final Answer: 10<br><br>*Student solution*: Step 1: Carlos memorized 18 digits of pi. Step 2: Mina memorized 24 digits of pi. Step 3: Sam memorized 6 more digits of pi than Carlos, so he memorized 24 digits of pi (18 + 6 = 24). Final Answer: 24 | T: (probing) You outlined that Carlos memorized 18 digits of pi - where did you get this figure from?<br><br>S: I got this figure by taking the number of digits that Mina memorized (24) and subtracting the number of digits that Carlos memorized more than Mina (6). So 24 - 6 = 18.<br><br>T: (probing) If the question says that Mina memorised six times as many digits of pi as Carlos memorised, and she memorised 24, what do you think that means for the digits of pi that Carlow memorised?<br><br>S: It means that Carlos memorized 4 digits of pi (24/6 = 4).<br><br>T: (focus) That's brilliant - really well done. So, if that means that Carlos memorised 4 digits of pi, and the question says that Sam memorised six more digits of pi than Carlos, how many digits of pi has Sam memorised?<br><br>S: Sam has memorized 10 digits of pi (4 + 6 = 10).<br><br>T: (focus) That's absolutely perfect! So, based on that, how many digits of pi did Carlos, Mina and Sam each memorise?<br><br>S: Carlos memorized 4 digits of pi, Mina memorized 24 digits of pi, and Sam memorized 10 digits of pi.<br><br>T: (focus) Well done Luca! You got it right! |
| | |

Table 6: Examples of conversations. $T$ refers to a teacher utterance, $S$ refers to a student utterance. Bold text is information for the reader indicating error categories.