

# Intent Matters: Enhancing AI Tutoring with Fine-Grained Pedagogical Intent Annotation

Kseniia Petukhova, Ekaterina Kochmar

Mohamed bin Zayed University of Artificial Intelligence  
{kseniia.petukhova, ekaterina.kochmar}@mbzuai.ac.ae

## Abstract

Large language models (LLMs) hold great promise for educational applications, particularly in intelligent tutoring systems. However, effective tutoring requires alignment with pedagogical strategies – something current LLMs lack without task-specific adaptation. In this work, we explore whether fine-grained annotation of teacher intents can improve the quality of LLM-generated tutoring responses. We focus on MathDial, a dialog dataset for math instruction, and apply an automated annotation framework to re-annotate a portion of the dataset using a detailed taxonomy of eleven pedagogical intents. We then fine-tune an LLM using these new annotations and compare its performance to models trained on the original four-category taxonomy. Both automatic and qualitative evaluations show that the fine-grained model produces more pedagogically aligned and effective responses. Our findings highlight the value of intent specificity for controlled text generation in educational settings, and we release our annotated data and code to facilitate further research: <https://github.com/Kpetyxova/autoTree/tree/main/mathdial>

## 1 Introduction

Human tutoring is a cornerstone of educational development, playing a vital role in empowering learners and fostering societal progress. One-on-one tutoring has long been recognized as highly effective (Bloom, 1984); however, its widespread implementation is constrained by the limited availability of qualified tutors. Recent advancements in LLMs have shown great promise in educational contexts (Wang et al., 2024; Gan et al., 2023), leading to the emergence of LLM-powered intelligent tutoring systems (ITS) (Pal Chowdhury et al., 2024; Liu et al., 2024) and the use of LLMs as tutors via advanced prompting strategies (Denny et al., 2024; Mollick and Mollick, 2024). These AI tutors serve

a range of educational objectives (Wollny et al., 2021), with one of the most prominent being the remediation of student mistakes and confusion – an area that continues to drive the development of AI tutoring systems (Macina et al., 2023; Wang et al., 2023).

While LLMs do well both at generating human-like conversations and at addressing various reasoning tasks, such as commonsense reasoning and basic mathematical reasoning (Achiam et al., 2023; Kojima et al., 2022; Laskar et al., 2023; Yang et al., 2024), they cannot be directly deployed in educational systems without significant adaptation. Effective tutoring requires more than fluent conversation – it involves guiding learners to discover answers on their own. Rather than simply providing solutions, a good tutor employs strategies such as giving hints, asking questions in a Socratic dialog (Carey and Mullan, 2004), and encouraging active problem-solving. As such, LLM-based tutors should ideally align with human tutoring strategies (Nye et al., 2014) and active learning practices shown to enhance student outcomes (Freeman et al., 2014).

In order to have such models, we need dialog tutoring datasets. MathDial (Macina et al., 2023) is one such dataset, comprising tutor-student dialogs centered around math reasoning tasks. Each teacher utterance is labeled with one of four pedagogical move types from Macina et al. (2023): *Focus* (guiding task progress), *Probing* (encouraging conceptual exploration), *Telling* (providing help when students are stuck), or *Generic* (non-pedagogical conversational turns). These annotations were provided by teachers during data collection to better scaffold student learning. While this four-category taxonomy offers a helpful high-level structure, it lacks the fine-grained detail needed for advanced applications such as controlled response generation, pedagogical analysis, and behavior modeling in AI tutors. At the same time,

finer-grained annotations may enable better interpretability, improved pedagogical alignment, and greater flexibility in guiding student learning experiences.

Although MathDial’s original taxonomy includes only four broad categories, the authors also provide an expanded set of eleven fine-grained intents, which could offer greater control and variety in AI-generated tutoring responses. Building on this, in this work, we apply a fully automated framework for conversational discourse annotation (Petukhova and Kochmar, 2025) to **re-annotate a portion of the MathDial dataset using the finer-grained eleven-intent taxonomy**. This annotation framework uses LLMs to automatically construct a decision tree from the taxonomy and use it to label utterances, providing a scalable alternative to manual annotation. This approach has demonstrated superior performance compared to crowdworkers in annotating dialog with speech functions taxonomy (Eggins and Slade, 2004).

**Our goal in this work is to assess whether such more detailed annotations can improve the quality of LLM-based tutoring through fine-tuning models on both the original and re-annotated data.** Specifically, we fine-tune Mistral-7B-Instruct on the original coarse-grained as well as the new fine-grained annotation, and compare the generated tutor responses using automatic metrics and human evaluation. Our results demonstrate that **the fine-grained model produces more pedagogically aligned and effective responses**. To facilitate further research and development, we release a public repository containing both the code and the re-annotated dataset.<sup>1</sup>

## 2 Background & Related work

### 2.1 The MathDial Dataset

We build on the foundational work of Macina et al. (2023), whose dataset provide an invaluable basis for advancing pedagogically aligned dialog systems. MathDial is a large-scale, high-quality dialog tutoring dataset focused on multi-step math reasoning problems. Unlike previous datasets that suffer from low pedagogical quality, small size, or lack of grounding, MathDial provides rich annotations grounded in realistic student confusions and pedagogical strategies. The authors introduce a novel semi-synthetic data collection framework

<sup>1</sup>Available at <https://github.com/Kpetyxova/autoTree/tree/main/mathdial>.

that pairs expert human teachers with LLMs simulating students and their errors, enabling scalable and controlled creation of educational dialogs that closely mimic authentic tutoring scenarios. This approach effectively addresses privacy concerns and quality issues associated with crowdsourcing or classroom recordings.

The authors’ methodology consists of a Wizard-of-Oz-inspired framework (Kelley, 1984), where expert teachers engage in one-on-one tutoring dialogs with LLMs acting as students. These student models are carefully prompted with student profiles and frequently occurring conceptual errors generated using temperature sampling over diverse reasoning paths produced by LLMs. The math word problems (MWP) used are sourced from GSM8K (Cobbe et al., 2021). Teachers are instructed to scaffold student understanding using a taxonomy of four pedagogical moves: *Focus*, *Probing*, *Telling*, and *Generic*, with additional fine-grained intents (see Table 1).

Crucially, before writing a response, teachers must annotate the pedagogical move being employed, encouraging more intentional strategy use. The dialogs are also grounded in metadata, including the specific confusion, full problem, step-by-step solutions, and whether the confusion was resolved, thus offering rich signals for training AI tutors.

Empirical evaluation demonstrates that models fine-tuned on MathDial significantly outperform both zero-shot and instruction-tuned larger LLMs like ChatGPT in terms of correctness and equitable tutoring (Macina et al., 2023). Notably, fine-tuned open-source models achieved similar rates of student problem-solving success while reducing the incidence of “telling” – prematurely giving away solutions. Human evaluations confirmed that these fine-tuned models were more coherent, correct, and pedagogically effective than large prompted models.

### 2.2 Annotation Framework

While manual discourse annotation is costly and time-consuming, advances in LLM-based annotation present a promising alternative with demonstrated improvements in speed, consistency, and cost-effectiveness (Gilardi et al., 2023; Hao et al., 2024). Petukhova and Kochmar (2025) have recently proposed an open-source pipeline for fully automated discourse annotation using LLMs. Specifically, this pipeline automates the construc-

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

Table 1: Teacher moves with examples of utterances and their intents from MathDial (Macina et al., 2023).

tion of hierarchical tree annotation schemes and the annotation of utterances within dialogs, making it a promising and scalable approach for enriching the MathDial dataset with more detailed teacher intent annotations.

Petukhova and Kochmar (2025) explore multiple configurations for tree construction and annotation, including binary and non-binary structures, frequency-based grouping, and optimal split strategies, and report that the frequency-guided optimal split selection using GPT-4o outperforms crowdworkers on dialog annotation tasks based on the taxonomy of speech functions (Eggin and Slade, 2004), while reducing total annotation time from over 30 hours to under 1.5 hours. Therefore, in our work, we adopt this configuration using the publicly available implementation.<sup>2</sup>

### 2.3 Controlled Generation

Controlled text generation (CTG) aims to direct language models to produce outputs that adhere to specific attributes or constraints, such as sentiment, style, or intent. A prevalent method in CTG involves fine-tuning models with prompts that include explicit intent labels, enabling the generation of text aligned with desired behaviors (Liang et al., 2024).

Instruction fine-tuning has emerged as an effective strategy for this purpose. By training models on datasets where prompts are augmented with natural language instructions or intent labels, models learn to condition their outputs accordingly. For instance, the InstructCTG framework demonstrates how conditioning on natural language descriptions and demonstrations of constraints allows models to generate text that satisfies various requirements without altering the decoding process (Zhou et al., 2023).

<sup>2</sup><https://github.com/Kpetyxova/autoTree>

This approach is particularly beneficial in educational contexts, where aligning generated content with pedagogical strategies is crucial. By fine-tuning models with prompts that specify instructional intents, AI tutors can provide more effective and tailored support to learners (Jia et al., 2025).

## 3 Re-annotating the MathDial Dataset

### 3.1 Tree Creation

To construct a tree for the extended taxonomy proposed in MathDial, we used the best framework configuration from Petukhova and Kochmar (2025) – frequency-guided optimal split selection and backtracking. This method iteratively selects among the candidate splits by scoring them and choosing the highest-ranked one, with backtracking employed if a viable partition cannot be formed. Additionally, the approach biases the tree construction toward more frequent classes, making them quicker to reach and producing trees that better reflect real-world class distributions. The tree was generated based on eleven intent names and their corresponding examples (see Table 1). The resulting tree, presented in textual form in Figure 1, has a depth of two with five branches emerging from the root node.

Interestingly, most intents are grouped according to high-level categories defined in Macina et al. (2023), except for the *Probing* intents, which are split into two separate groups: (1) *Asking for Explanation* and *Seeking Self-Correction*, and (2) *Perturbing the Question* and *Seeking World Knowledge*. While this split was not predefined, it is interpretable: the first group centers on prompting students to reflect on and assess their reasoning, whereas the second group encourages them to explore broader or external concepts beyond the immediate problem.

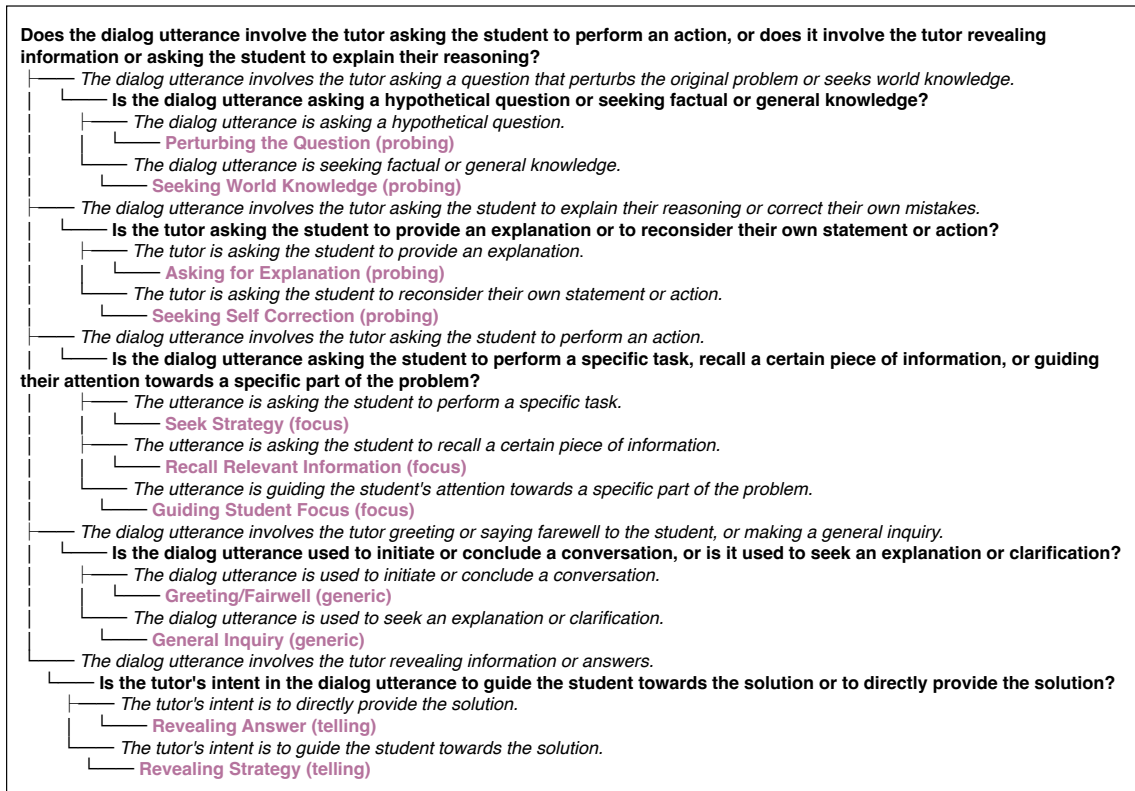


Figure 1: Tree created for the extended taxonomy of the MathDial dataset using the framework from [Petukhova and Kochmar \(2025\)](#). Questions corresponding to tree nodes are in **bold**, possible answers that represent branches are in *italics*, and leaf nodes, representing the eleven intents, are in **purple bold**.

### 3.2 Annotation

**Data Preprocessing** Out of 2,861 dialogs, we randomly selected 500 dialogs for training, 100 for validation, and 100 for testing.

An example of the original tutor intent annotation in MathDial is shown in Figure 2. A single label is applied to each teacher utterance in the original annotation, which, while effective for high-level analysis, may limit flexibility in downstream applications requiring finer-grained control. For instance, an utterance [I see.]<sub>1</sub> [But we’re dealing with individual pies here, rather than slices.]<sub>2</sub> [If you had a birthday cake, and lots of guests at your party, you couldn’t just keep producing slices of cake.]<sub>3</sub> [Can you think of another way to figure out how everyone gets a piece?]<sub>4</sub> in MathDial is annotated as *Probing*. However, this utterance comprises several discourse units with distinct functions: segment [1] appears to be *Generic*, segment [2] aligns with *Focus* (specifically, *Guiding Student Focus*) as it redirects the student’s attention, segment [3] fits the *Probing* category, and segment [4] corresponds to *Focus* (*Seek Strategy*) because

it prompts the student to think of an alternative solution.

In contrast, in other cases, the assigned label appears to follow the final part of the utterance. For example, the utterance [But there are always 4 slices in a shepherd’s pie, so using the total number of slices might not be helpful.]<sub>1</sub> [Are there any other quantities you could use to divide by the slices in the pie?]<sub>2</sub> is labeled as *Focus*. Here, while *Focus* applies to the second sentence [2], it would be more appropriate to label the first sentence [1] as *Telling*. This inconsistency – where labels are sometimes based on the first segment and other times on the last – underscores the potential benefits of a more fine-grained and consistent annotation approach for certain downstream tasks.

Ideally, annotation should be performed at the elementary discourse units (EDUs) level rather than entire utterances. EDUs are segments of text that typically correspond to clauses ([Jurafsky and Martin](#)). Therefore, in this work, we preprocess the data by first splitting teacher utterances into EDUs.

Since no state-of-the-art method currently ex-



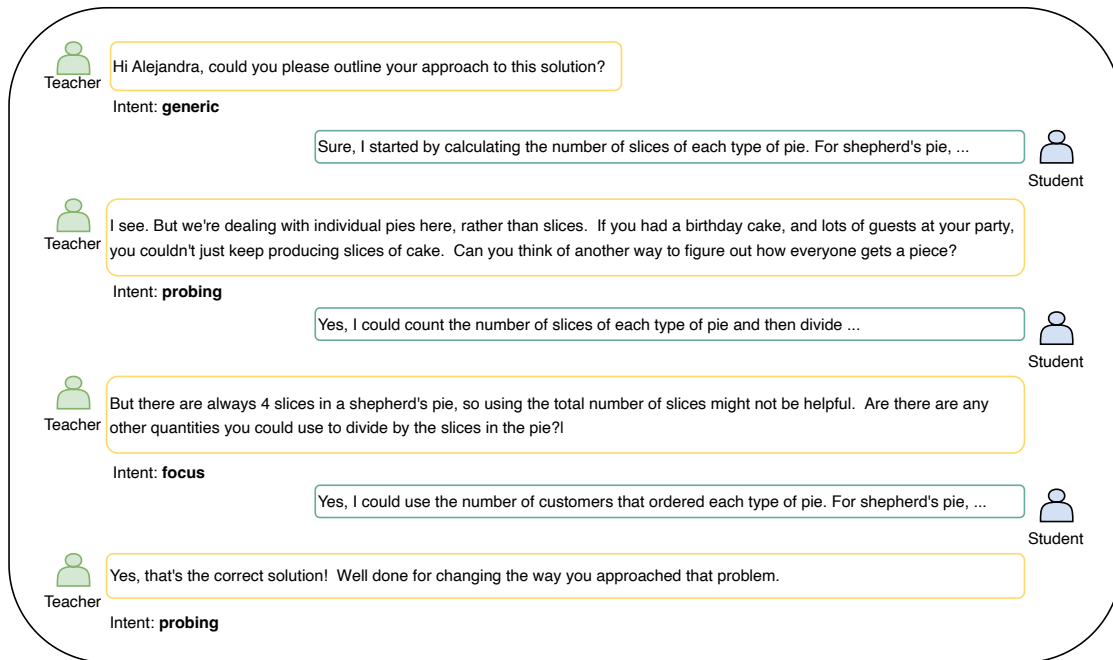


Figure 2: An example of teacher utterances and their annotated intents from MathDial.

ists for automatically dividing the text into EDUs, we use the following strategy: (1) Punctuation Removal: first, we remove all punctuation from the utterances; (2) Punctuation Restoration: next, we restore the punctuation using a model trained for this task;<sup>3</sup> (3) Comparison and Segmentation: finally, we compare the original utterance with the punctuation-restored version. If the restored punctuation replaces a comma in the original text with a period, question mark, or exclamation mark, we split the utterance at that comma, thereby creating separate EDUs. By default, we also split different sentences into separate EDUs. Each EDU that resulted from the original utterance through this process inherits the original label assigned to the full utterance in MathDial (i.e., one of the four high-level categories).

After the data is split into EDUs, the number of resulting teacher utterances in the **train** split is **5,174**. The **validation** and **test** sets are similarly segmented into EDUs and limited to **100** teacher utterances each.

**Annotation** Using the generated tree, a GPT-4o-based annotation pipeline from [Petukhova and Kochmar \(2025\)](#) is applied. Since the tree’s structure aligns with the hierarchical intent relationships

<sup>3</sup><https://huggingface.co/oliverguhr/fullstop-punctuation-multilang-large>

proposed by the authors of MathDial, we can reasonably expect that annotation based on this tree will reflect those relationships. For instance, if the annotation using the tree assigns the label *Perturbing the Question*, the original annotation should correspondingly contain *Probing*, and so on. Based on this alignment, we can evaluate the annotation quality, at least in terms of consistency with the original higher-level annotations.

Table 2 presents weighted precision ( $P_w$ ), recall ( $R_w$ ), and F1 ( $F1_w$ ) as well as macro F1 (F1) scores when comparing lower-level intent annotations on the training set to the original high-level teacher move categories. The low scores are expected, given that the original teacher utterances were split into EDUs while retaining the same label. As discussed earlier, different EDUs within the same utterance would often be of distinct types, which was not accounted for in the original annotation in MathDial.

$P_w$	$R_w$	$F1_w$	F1
0.40	0.38	0.36	0.27

Table 2: Evaluation of 11-label annotation on the training set, comparing the new alignment with the original 4-label annotation from MathDial, using the annotation framework from [Petukhova and Kochmar \(2025\)](#).

Among the 5,174 teacher utterances, 1,319 remained unchanged from the original dataset, as

they originally consisted of a single EDU. Annotation results for these utterances are presented in Table 3. While these metrics are higher than those in Table 2, they still indicate relatively poor performance.

$P_w$	$R_w$	$F1_w$	$F1$
0.48	0.45	0.43	0.31

Table 3: Evaluation of 11-label annotation on the training set utterances that remained unchanged (i.e., originally consisted of a single EDU), comparing the new alignment with the original 4-label annotation from MathDial, using the annotation framework from Petukhova and Kochmar (2025).

However, a manual analysis revealed significant inconsistencies in the original annotation. Consider the following illustrative examples:

- A student initially identifies 14 as the correct final answer to the task. However, during the discussion, the student incorrectly restates the final solution as  $10 + 10 + 4 = 24$ . The teacher responds, *Is that 14?* — referring back to the earlier moment when 14 was correctly identified as the expected answer (see the full dialog in Appendix A). The tree-based annotation classifies this teacher utterance as *Seeking Self-Correction*, corresponding to *Probing*. However, in the original dataset, it is labeled as *Telling*, which we believe is not accurate.
- The tutor says, *You need to add brackets to (8-2) and remember the order of operations.* The student responds, *Yes, I understand now. The correct equation should be  $6 + (8 + 8) - 2 = 22$  new books.* The teacher replies, *No, I said it's 8-2, not 8+8.* Although the tree-based annotation assigns this utterance to *Revealing Answer (Telling)*, the original annotation labels it as *Generic*, possibly reflecting a different interpretation or contextual judgment.
- A student states,  $6 + 8 + (8 - 2) = 22$ . The teacher responds, *Please explain how you got 22.* The tree-based annotation categorizes this utterance as *Asking for Explanation*, which corresponds to *Probing*. However, in the original dataset, it is labeled as *Generic*, which does not align well with the intent of the utterance.

Given the prevalence of such unclear or ambiguous cases in the original annotation of the dataset,

we cannot conclude that the tree-based annotation is inaccurate. Instead, these inconsistencies in the original annotation suggest that the discrepancies in the evaluation metrics may be due, at least in part, to ambiguities in the original dataset.

The distribution of the eleven predicted intents across all dataset splits (train, validation, and test) is shown in Figure 3.

## 4 Controlled Generation

To demonstrate the benefits of an extended taxonomy with annotations collected using the framework from Petukhova and Kochmar (2025), we fine-tune an LLM to predict the next teacher utterance. The model is trained using the math task description, its gold solution, the student’s solution, the dialog history, and the teacher’s next utterance intent as predicted by the annotation framework.

Additionally, we fine-tune a second version of the same model using the original four-intent annotation. We then compare the performance of these two fine-tuned models with each other, as well as with the same LLM in its zero-shot setting.

**Model** We use Mistral-7B-Instruct as the base model for fine-tuning, specifically its 4-bit quantized version from Hugging Face.<sup>4</sup> The maximum sequence length is set to 1,600. We fine-tune the model using QLoRA (Quantized Low-Rank Adaptation) (Hu et al., 2022), a parameter-efficient method that applies low-rank adapters with quantization to reduce memory and compute costs. We use a rank of  $r = 32$  and scaling factor  $\alpha = 32$ . Fine-tuning is conducted for one epoch with a learning rate of  $2e^{-5}$ , batch size 8, and gradient accumulation of 4. We employ the AdamW optimizer (Loshchilov and Hutter, 2019), linear scheduling with a warmup (0.1), weight decay of 0.1, and evaluate every 50 steps using SACREBLEU (Post, 2018).

**Data Preprocessing** We convert the annotated samples into pairs of prompts and gold outputs, where each prompt consists of an instruction, the math task, the gold solution for the task, the student’s solution, the dialog history, and the intent of the following teacher utterance (which is available from the annotated data). While the intent is available as an annotation during both training and evaluation – since we have access to the gold next

<sup>4</sup><https://huggingface.co/unsloth/mistral-7b-instruct-v0.3-bnb-4bit>

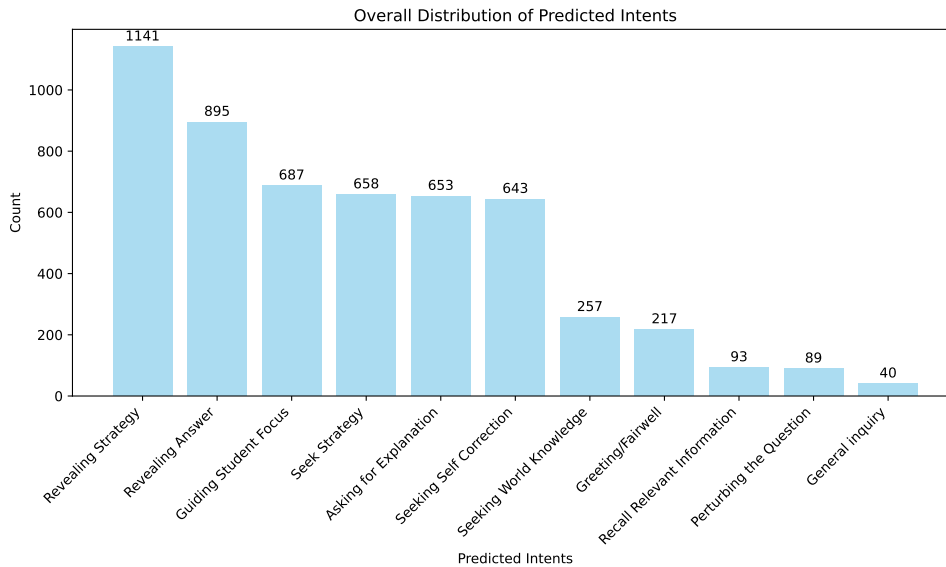


Figure 3: Overall distribution of the eleven predicted intents in the re-annotated dataset.

teacher utterance and can classify its intent – for real-world applications this intent would need to be predicted by a separate model as part of a controlled generation pipeline. The prompt template is shown in Appendix B.

**Evaluation** We conduct an automatic evaluation of generated outputs using reference-based metrics, including CHRF++ (character n-gram  $F$ -score) (Popović, 2017), SACREBLEU (a weighted geometric mean of  $n$ -gram precision scores), and ROUGE-1, ROUGE-2, and ROUGE-L (recall-oriented measures of  $n$ -gram overlap) (Lin, 2004). In addition, we conduct a small-scale human evaluation.

**Results** Table 4 presents the generation results for both zero-shot and fine-tuning settings, comparing two annotation schemes: the original four teacher intents provided in the MathDial dataset and the extended set of eleven intents. As expected, the fine-tuned LLM outperforms the zero-shot baseline, and the model trained on the more fine-grained, eleven-intent annotation consistently achieves higher scores across all metrics.

In addition to automated metrics, we conducted a human evaluation with four annotators, each holding at least a Master’s degree in Natural Language Processing. We randomly selected seven dialogs from the test set, resulting in 30 response pairs – one from the model fine-tuned on four intents and one from the model fine-tuned on eleven intents. Each annotator was shown these pairs and asked to decide which response was better or whether

both were equally good or poor (see Figure 4). Based on majority voting, responses from the FT-11 model were preferred in 56.7% of cases.<sup>5</sup> The inter-annotator agreement, measured using Fleiss’ Kappa, is  $\kappa = 0.33$ , indicating fair agreement.

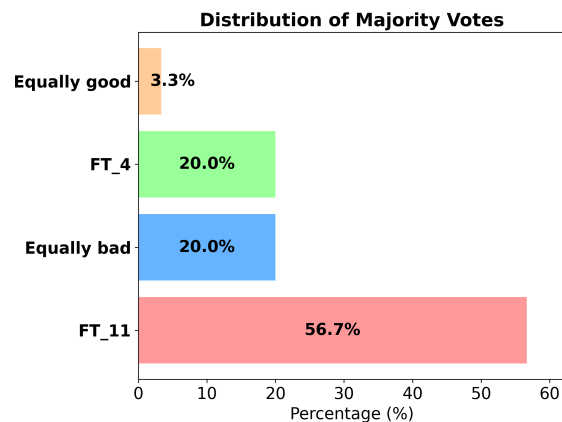


Figure 4: Results of human evaluation. Four annotators were asked to choose the better response or indicate if both were equally good or bad. Final decisions were determined via a majority vote.

Manual analysis (see Figure 5) indicates that the responses generated by the model fine-tuned on eleven intents (FT-11) are superior to those from the model fine-tuned on four intents (FT-4), based on the following observations:

- FT-11 consistently demonstrates a deeper understanding of conversational strategies, such

<sup>5</sup>There were no ties in the majority votes – each example received a clear decision.

Configuration	CHRF++	SACREBLEU	ROUGE-1	ROUGE-2	ROUGE-L
Zero-Shot, 4 intents	16.50	0.93	8.93	2.19	7.10
Zero-Shot, 11 intents	17.11	0.73	8.72	1.95	6.87
Fine-Tuning, 4 intents	16.82	2.67	17.13	5.61	15.95
Fine-Tuning, 11 intents	<b>18.06</b>	<b>4.59</b>	<b>20.73</b>	<b>7.39</b>	<b>19.28</b>

Table 4: Evaluation of controlled generation on the test set from MathDial in a zero-shot setting and with fine-tuned Mistral, comparing using the original four intents from MathDial with eleven intents annotated using the framework from [Petukhova and Kochmar \(2025\)](#).

as using more effective questioning techniques. For instance, when addressing the incorrect prom couples calculation (see the first example in Figure 5), FT-11 explicitly prompts the student to reconsider the original conditions (*So, if we know that there were 123 students at the prom, how many couples were there?*). In contrast, FT-4 merely restates the incorrect scenario (*So, if we have 120 couples, how many students attended the prom?*), which is less effective in guiding the student to realize their mistake.

- FT-11 more directly addresses student misconceptions. In the second example in Figure 5, FT-11 directly questions the student’s arbitrary assumption (*How did you get 100 cows?*), aligning closely with the teacher’s gold standard (*Claire, why did you assume that the farmer had 100 cows?*). FT-4 is less focused, requesting the student to explain calculations instead of addressing the root cause of misunderstanding.
- FT-11 responses tend to be concise yet relevant, prompting students to reflect critically on their reasoning rather than reiterating previous statements. For example, in the third scenario in Figure 5, FT-11 succinctly acknowledges correctness (*Correct.*), aligning well with the actual teacher response (*That’s right.*), while FT-4 unnecessarily repeats previous questions, demonstrating less effective dialog management.

## 5 Conclusions

In this work, we examined the impact of fine-grained annotation on controlled response generation in the MathDial dataset. By expanding the original taxonomy of teacher moves from four broad categories to eleven more specific intents and using the framework for automated tree creation and annotation from [Petukhova and Kochmar](#)

(2025), we demonstrated that this approach enhances the performance of a fine-tuned LLM in generating meaningful responses.

The results confirm that fine-tuning on a dataset with high-granularity labels leads to better alignment with expected teacher responses, outperforming both the zero-shot setting and fine-tuning on the original four-category annotation. This suggests that the specificity of intent labels is crucial for enhancing the model’s ability to generate targeted and effective tutoring responses.

Furthermore, the manual analysis of the annotation process revealed inconsistencies in the original dataset, highlighting the importance of high-quality, well-structured intent annotation. The tree-based annotation framework proposed in [Petukhova and Kochmar \(2025\)](#) provides a systematic way to refine such datasets, making them more suitable for training controllable generation models.

## Limitations

Due to resource constraints, we re-annotate only a subset of the MathDial dataset and utilize a small language model for fine-tuning.

These limitations suggest several directions for future work: (1) re-annotating the entire dataset to enable training of higher-quality models, and (2) exploring larger open-source models for improved fine-tuning performance.

In addition, we acknowledge that our conclusions are based on the use of automated metrics and small-scale human evaluation. In order to demonstrate the real-world impact of these findings and assess the pedagogical value of the generated tutor interventions, future work should consider verifying these conclusions with actual teachers and students.

## Ethical Considerations

As this work is exploratory and the outputs of the models used in this research have not been tested with real students, we do not anticipate any significant risks associated with this work or the use



Task	Student's Incorrect Solution	Dialog
Chalktown High School had their prom last weekend. There were 123 students who attended. If 3 students attended on their own, how many couples came to the prom?	Let's call the number of couples that attended the prom "c". We know that there were 123 students total, and that 3 of them attended on their own. This means that the rest of the students attended as couples. So we can set up the equation: couples + 3 = 123 Subtracting 3 from both sides: couples = 120 So there were 120 couples that attended the prom. 120	<p><b>Teacher:</b> Hey Heidi, do you know what a couple is?</p> <p><b>Student:</b> Yes, a couple is two people who are together, usually a romantic couple.</p> <p><b>Teacher:</b> So if we had 120 couples attending the prom, how many students would that be in total?</p> <p><b>Student:</b> 120 couples would be 240 students in total, since each couple is made up of two people.</p> <p><b>Teacher:</b> Great.</p> <p><b>Teacher:</b> would you like to check your solution again, bearing in mind what youve just told me, please?</p> <p><b>Student:</b> Sure. So if we have 120 couples, that would mean there are 240 students in total. We know that 3 students attended on their own, so if we subtract 3 from 240, we get 237. So there were 120 couples and 237 students in total at the prom.</p> <p><b>Teacher:</b> We need to read the question again.</p> <p><b>Teacher:</b> the question tells us how many people were at the prom.</p> <p><b>Next Teacher Utterance:</b>  <b>Gold:</b> lets start over from the beginning and ill help you if needed.  <b>FT-4 (focus):</b> So, if we have 120 couples, how many students attended the prom?  <b>FT-11 (Seek Strategy):</b> So, if we know that there were 123 students at the prom, how many couples were there?</p>
40% of a farmer's cattle are males. The rest are females. If a female cow produces 2 gallons of milk a day, how much milk will the farmer get a day if he has 50 male cows?	If 40% of the cattle are males, then 100% - 40% = 60% are females. Let's say the farmer has a total of 100 cows, then 40% of 100 = 40 cows are males. So 60 cows are females. If each female cow produces 2 gallons of milk a day, then 60 x 2 = 120 gallons of milk will be produced a day. Since the farmer has 50 male cows, they don't produce milk, so the total milk produced a day is 120 gallons.	<p><b>Teacher:</b> Hi, could you please walk me through your solution?</p> <p><b>Student:</b> Sure. I started by figuring out that 40% of the cattle were males, so 100% - 40% = 60% were females. Then I assumed that the farmer had a total of 100 cows, so 40% of 100 = 40 cows were males and 60 cows were females. I then multiplied the number of female cows (60) by 2 gallons of milk a day to get 120 gallons of milk produced a day. Since the farmer only has 50 male cows, they don't produce any milk, so the total milk produced a day is 120 gallons.</p> <p><b>Next Teacher Utterance:</b>  <b>Gold:</b> Claire, why did you assume that the farmer had 100 cows?  <b>FT-4 (generic):</b> Hi, could you please explain how you got 120 gallons of milk from 60 cows?  <b>FT-11 (Seek Strategy):</b> How did you get 100 cows?</p>
It takes 3 men an hour to complete a job. If they are hired to work on 5 such similar jobs by a company that pays each of them \$10 per hour, calculate the total amount of money the three will earn when they complete the jobs?	It takes 3 men an hour to complete a job, so it will take them 3 x 5 = 15 hours to complete the 5 jobs. If they are paid \$10 per hour, the total amount of money each of them will earn when they complete the jobs is \$10 x 15 = \$150. Therefore, the total amount of money the three will earn when they complete the jobs is \$150 x 3 = \$450.	<p><b>Teacher:</b> Hi Brenda.</p> <p><b>Teacher:</b> lets take this step by step.</p> <p><b>Teacher:</b> How much would one man earn for one job?</p> <p><b>Student:</b> Hi, one man would earn \$10 for one job.</p> <p><b>Teacher:</b> How many jobs are there?</p> <p><b>Student:</b> There are 5 jobs.</p> <p><b>Teacher:</b> How much would one man doing the 5 jobs earn?</p> <p><b>Student:</b> One man doing the 5 jobs would earn \$10 x 5 = \$50.</p> <p><b>Teacher:</b> Great, How many men are there?</p> <p><b>Student:</b> There are 3 men.</p> <p><b>Teacher:</b> So each man is paid 50 total for the 5 jobs and there are 3 men.</p> <p><b>Teacher:</b> so how much is the total for the 3 men?</p> <p><b>Student:</b> The total for the 3 men is \$50 x 3 = \$150.</p> <p><b>Next Teacher Utterance:</b>  <b>Gold:</b> Thats right.  <b>FT-4 (generic):</b> So how much is the total for the 3 men?  <b>FT-11 (Revealing Answer):</b> Correct.</p>

Figure 5: Examples of responses generated by a model fine-tuned on the original four-intent annotation set (FT-4) and by a model fine-tuned on data annotated with eleven intents using the framework from Petukhova and Kochmar (2025) (FT-11), alongside actual responses from teachers (Gold).

of the re-annotated dataset. At the same time, we acknowledge that this work uses LLMs, and such models may present risks when applied in real-life educational scenarios, as they may generate outputs that, despite being plausible, are factually inaccurate or nonsensical, which in turn may lead to misguided decision making and propagation of biases. While we do not foresee any immediate risks associated with the research presented in our paper, if future work based on this research ap-

plies presented approaches to real-life scenarios, appropriate safeguards should be applied.

## Acknowledgments

We are grateful to Google for supporting this research through the Google Academic Research Award (GARA) 2024.

## References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Benjamin S Bloom. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher*, 13(6):4–16.
- Timothy A Carey and Richard J Mullan. 2004. What is Socratic questioning? *Psychotherapy: theory, research, practice, training*, 41(3):217.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems, 2021. URL <https://arxiv.org/abs/2110.14168>, 9.
- Paul Denny, Sumit Gulwani, Neil T Heffernan, Tanja Käser, Steven Moore, Anna N Rafferty, and Adish Singla. 2024. Generative AI for education (GAIED): Advances, opportunities, and challenges. *arXiv preprint arXiv:2402.01580*.
- Suzanne Eiggins and Diana Slade. 2004. *Analysing casual conversation*. Equinox Publishing Ltd.
- 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.
- Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Jerry Chun-Wei Lin. 2023. Large language models in education: Vision and opportunities. In *2023 IEEE international conference on big data (BigData)*, pages 4776–4785. IEEE.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
- Jing Hao, Yuxiang Zhao, Song Chen, Yanpeng Sun, Qiang Chen, Gang Zhang, Kun Yao, Errui Ding, and Jingdong Wang. 2024. Fullanno: A data engine for enhancing image comprehension of mllms. *arXiv preprint arXiv:2409.13540*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Linzhao Jia, Changyong Qi, Yuang Wei, Han Sun, and Xiaozhe Yang. 2025. Fine-Tuning Large Language Models for Educational Support: Leveraging Gagne’s Nine Events of Instruction for Lesson Planning. *arXiv preprint arXiv:2503.09276*.
- Daniel Jurafsky and James H Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*.
- 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.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. *arXiv preprint arXiv:2305.18486*.
- Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and 1 others. 2024. Controllable text generation for large language models: A survey. *arXiv preprint arXiv:2408.12599*.
- Chin-Yew Lin. 2004. **ROUGE: A package for automatic evaluation of summaries**. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Rongxin Liu, Carter Zenke, Charlie Liu, Andrew Holmes, Patrick Thornton, and David J Malan. 2024. Teaching CS50 with AI: leveraging generative artificial intelligence in computer science education. In *Proceedings of the 55th ACM technical symposium on computer science education V. 1*, pages 750–756.
- Ilya Loshchilov and Frank Hutter. 2019. **Decoupled Weight Decay Regularization**. In *International Conference on Learning Representations*.
- Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. *arXiv preprint arXiv:2305.14536*.
- Ethan Mollick and Lilach Mollick. 2024. Instructors as innovators: A future-focused approach to new AI learning opportunities, with prompts. *arXiv preprint arXiv:2407.05181*.
- 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.
- Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. 2024. Autotutor meets large language models: A language model tutor with rich pedagogy and guardrails. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 5–15.

- Kseniia Petukhova and Ekaterina Kochmar. 2025. A Fully Automated Pipeline for Conversational Discourse Annotation: Tree Scheme Generation and Labeling with Large Language Models. *arXiv preprint arXiv:2504.08961*.
- Maja Popović. 2017. *chrF++: words helping character n-grams*. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. 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.
- Rose E Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2023. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. *arXiv preprint arXiv:2310.10648*.
- Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*.
- Sebastian Wollny, Jan Schneider, Daniele Di Mitri, Joshua Weidlich, Marc Rittberger, and Hendrik Drachslar. 2021. Are we there yet?-a systematic literature review on chatbots in education. *Frontiers in artificial intelligence*, 4:654924.
- Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.
- Wangchunshu Zhou, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, and Mrinmaya Sachan. 2023. Controlled text generation with natural language instructions. In *International Conference on Machine Learning*, pages 42602–42613. PMLR.

## A Output of the Tree Creation Step

**Task:** Harry decided to buy some balloons for his friend's birthday party. One balloon costs \$0.5, and a pack of 10 balloons is cheaper and costs only \$3. Finally, Harry wants to bring to the party exactly 14 balloons. How much did he need to pay for them?

**Correct Solution:** To buy 14 balloons, Harry should buy one pack of 10, and  $14 - 10 = 4$  separate balloons.

The price of 4 balloons is  $4 * 0.5 = \$2$ .

In total, the pack and separate balloons would cost Harry in total  $3 + 2 = \$5$ .

5

**Student's incorrect answer:** Harry needs to buy 2 packs of 10 balloons, which will give him 20 balloons in total. He will then need to buy 4 more individual balloons to get to 14.

2 packs of 10 balloons cost  $2 * \$3 = \$6$ .

4 individual balloons cost  $4 * \$0.5 = \$2$ .

Therefore, Harry needed to pay  $\$6 + \$2 = \$8$  for 14 balloons.

8

**Dialog:**

Teacher: (generic)Hi Michael, could you please walk me through your solution?

Student: Sure. I calculated that Harry needed to buy 2 packs of 10 balloons, which cost  $2 * \$3 = \$6$ . Then, he needed to buy 4 individual balloons, which cost  $4 * \$0.5 = \$2$ . So, the total cost was  $\$6 + \$2 = \$8$ .

Teacher: (focus)So can you please tell me why he needs two packs of 10 (which is 20) and another 4 on top? What's  $20 + 4$ ?

Student: Harry needs two packs of 10 balloons because each pack costs \$3 and buying two packs is cheaper than buying 20 individual balloons. Then, he needs to buy 4 more individual balloons to get to 14 balloons in total.  $20 + 4 = 24$ , so Harry has enough balloons for the party.

Teacher: (focus)Please could you tell me how many balloons Harry requires in the math problem.

Student: Harry needs 14 balloons for the party.

Teacher: (focus)So is that 24?EOM!Student: No, 14. He needs two packs of 10 balloons and 4 individual balloons, which gives him a total of 24 balloons, but he only needs 14.

Teacher: (focus)So how many packs of 10, plus single balloons does he really need to equal 14?

Student: Harry needs two packs of 10 balloons and 4 individual balloons to equal 14.

Teacher: (probing)so what's  $10+10+4$ ?

Student:  $10+10+4 = 24$ .

Teacher: (telling)Is that 14?

Student: No, 24 is not 14. Harry needs two packs of 10 balloons and 4 individual balloons, which gives him a total of 24 balloons, but he only needs 14.

Teacher: (telling)So Harry needs one pack of 10, which equals \$3. Plus he needs four separate balloons at \$0.5 each, can you tell me how much he needs to spend?

Student: Harry needs to spend \$3 for one pack of 10 balloons and \$2 for four individual balloons, so the total cost is  $\$3 + \$2 = \$5$ .

Teacher: (generic)Brilliant, can you see where you were duplicating the packs of 10

Student: Yes, I was calculating that Harry needed to buy two packs of 10 balloons, which cost  $2 * \$3 = \$6$ . Then, I was calculating that he needed to buy 4 individual balloons, which cost  $4 * \$0.5 = \$2$ . So, the total cost was  $\$6 + \$2 = \$8$ , which was incorrect because I was duplicating the packs of 10.

Figure 6: An example dialog from the MathDial dataset.



## B The Prompt Template used for Fine-tuning

```
[INST]
You are a math tutor. You will receive a task, a correct ("gold") solution, a student's
incorrect solution, and an ongoing dialog between the student and the tutor.
Your objective is to generate the next tutor response based on the given intent.

Task:
{task}

Gold Solution:
{gold_solution}

Student's Incorrect Solution:
{student_incorrect_solution}

Dialog:
{dialog}

Intent for the Next Tutor Utterance:
{intent}

[/INST]
### Tutor:
```

Figure 7: The prompt template used for fine-tuning on MathDialog.