"Keep up the good work!": Using Constraints in Zero-Shot Prompting to Generate Supportive Teacher Responses

E. Margaret Perkoff [1], Angela Ramirez[2], Sean Von Bayern[1],
Marilyn Walker[2], James Martin[1]
[1]University of Colorado Boulder, [2]University of California Santa Cruz
Correspondence: elpe8468@colorado.edu

Abstract

Educational dialogue systems have been used to support students and teachers for decades. Such systems rely on explicit pedagogically-motivated dialogue rules. With the ease of integrating large language models (LLMs) into dialogue systems, applications have been arising that directly use model responses without the use of human-written rules, raising concerns about their use in classroom settings. Here, we explore how to constrain LLM outputs to generate appropriate and supportive teacher-like responses. We present results comparing the effectiveness of different constraint variations in a zero-shot prompting setting on a large mathematics classroom corpus. Generated outputs are evaluated with human annotation for Fluency, Relevance, Helpfulness, and Adherence to the provided constraints. Including all constraints in the prompt led to the highest values for Fluency and Helpfulness, and the second highest value for Relevance. The annotation results also demonstrate that the prompts that result in the highest adherence to constraints do not necessarily indicate higher perceived scores for Fluency, Relevance, or Helpfulness. In a direct comparison, all of the non-baseline LLM responses were ranked higher than the actual teacher responses in the corpus over 50% of the time.

1 Introduction

Large language models (LLMs) have shown great promise across many applications including recommendation systems, social chatbots, writing code, and summarizing documents (Zhang et al., 2023). Many of these applications benefit from the generative capabilities of LLMs such as ChatGPT (Brown et al., 2020). However, when these models are deployed without further constraints in open-domain dialogue systems, they may generate outputs that do not adhere to the desired agent behavior (Kann et al., 2022). The risk of not adhering to desired

Teacher	How do you know number two is not a straight angle?
Student	Because a straight angle goes on, on both sides.
Teacher	It goes on forever. There's a line, and it goes on forever on both sides, absolutely. Of course, it's a ray so it only goes in one direction.

Table 1: A sample interaction between a teacher and a student from the NCTE Corpus (Demszky and Hill, 2023)

agent behavior is even higher when we consider the application of these models to the education domain (Williams et al., 2023).

Well before the advent of LLMs, educational dialogue systems have been used to provide support to students in online classes, act as 1:1 subject-specific tutors, and provide professional training to teachers and tutors (Kuhail et al., 2023). Regardless of the exact learning application, they are more beneficial to students when the systems themselves are based on the same pedagogical frameworks that a teacher would use (Järvelä and Hadwin, 2013). As such, many of these systems are built using extensive dialogue frameworks that dictate when, and how the teacher should intervene in a particular scenario. Frequently, they are designed such that a teacher can modify the exact responses to apply to a given classroom age, subject, and lesson-specific goals. Although this makes it possible to create highly relevant responses from the conversational agent, it also means that the agent will not generalize well to new situations. Furthermore, these agents are typically designed for interaction with a single student working in an online classroom.

The educational dialogue system that we present here is designed to provide support to students in a real classroom setting. Consider the interaction in

Table 1. In this case, the teacher confirms that the student's understanding is correct and elaborates on what has been said to avoid confusion in the definition. This ensures that they are still encouraging the student to elaborate on their reasoning while providing a fully correct assessment of the problem. This highlights the degree of nuance required when responding to a student in a classroom setting. We want to focus on how to constrain the output of an LLM in a manner consistent with how a teacher would respond. Constraints are surfaced to the LLM through strategic prompt engineering. The constraints are based on dialogue acts that capture speaker intentions from the preceding dialogue and a domain-specific dialogue policy. The dialogue policy consists of three conversational states identified in collaborative task conversations and corresponding actions that a supportive teacher would take in that particular scenario. We evaluate the effectiveness of imposing different forms of constraints on the NCTE Corpus (Demszky and Hill, 2023) of math classroom transcripts. The output from each of the prompt variations is assessed for Fluency, Relevance, Helpfulness, and adherence to the desired constraints. We also perform further annotations for overall best response and whether it was considered better than the teacher.

The dialogue system we present here provides a blueprint for how future researchers, and teachers themselves, can modify prompts to better engage with students in a classroom setting. This design is intended to echo the highly relevant and helpful nature of early rule-based education dialogue systems while allowing for more flexibility with LLMs to surface the response to the student. The major contributions of this work are as follows:

- As far as we are aware, this is the first application of dialogue-policy informed LLM response generation to the education domain.

- Adding more context-specific constraints (Dialogue States, Student and Teacher Dialogue Acts) can improve the perceived Helpfulness, Fluency and Relevance of responses produced by a conversational agent.

- LLMs can adhere to dialogue state constraints in up to 95% of samples

- Annotators rated the non-baseline LLM-based prompt variations to be better than the actual Teacher response in > 50% of samples

The results demonstrate the potential of LLM-based chatbots to interact with students in a helpful manner. There is much research to be done in exploring how to balance additional pedagogical constraints while maintaining a high degree of fluency in the responses. In future work, we intend to integrate these models with a speech-to-text interface to see how they perform in live classroom interactions.

2 Related Work

2.1 Dialogue Act Segmentation and Classification

To respond in a succinct manner, dialogue systems need to be able to differentiate different dialogue acts (DAs) such as statements, types of questions, and acknowledgements. There are different dialogue act schemas and datasets for dialogue act tagging such as: ViGGO a video game corpus tagged with dialogue acts meant for open domain systems (Juraska et al., 2019), MultiWoz a multi-domain and topic dataset meant for task-oriented dialogue systems (Budzianowski et al., 2018), Switchboard (SWDA) is a large multi-speaker dataset consisting of two-sided telephone calls (Stolcke et al., 2000a) and the AMI meeting corpus that is multi-modal corpus consisting of 100 hours of meeting recordings (Shang et al., 2018). ISO dialogue acts schema are mapped to other corpora such as SWDA, AMI, Maptask to then be used for training a SVM model on ISO tags (Mezza et al., 2018a; Thompson et al., 1993).

Previous work, has used this ISO dialogue act model off-the-shelf model to enrich TopicalChats with dialogue acts to then use the dialogue acts for response generation (Hedayatnia et al., 2020). LLMs such as gpt-3.5-turbo have been used for data augmentation and annotation of dialogue acts on outputs within the education domain (Shan et al., 2023). We expand on these papers by tagging dialogue acts using a combination of ISO tags and Switchboard feedback dialogue acts and by using using gpt-3.5-turbo on the NCTE dataset to produce a silver set of annotated dialogue acts as this corpus does not contain tags. Then we leverage this new tagged dataset as a constraint to provide the model for controllable response generation (Ramirez et al., 2023; Hedayatnia et al., 2020).

Dialogue act tagging and segmentation can be split into two separate tasks such as dialogue act classifiers (Stolcke et al., 2000a; Webb and Wilks,

2005), or a joint approach (Guz et al., 2010; Zhao and Kawahara, 2019, 2017). We take inspiration from joint models by combining the two tasks within the same prompt. The description used to prompt the model resembles annotation manuals that request annotators to consider both segmenting and tagging each segment with a dialogue act (Mezza et al., 2018a).

2.2 Pedagogical Conversational Agents

Pedagogical conversational agents are the subset of language models that can engage in dialogues to support learning. They vary greatly in terms of their role, their interaction style, and their functional purpose. Recent reviews of pedagogical conversational agents have found that they are frequently used as Teaching Agents in the context of online classroom settings, with the majority focused on Computer Science classroom courses (Kuhail et al., 2023). Although early research focused more on rule-based dialogue systems to power these agents, the surge in popularity of generative language models has led to more research examining the ability of language models to generate teacher-like responses. Tack and Piech (2022) proposed the AI Teacher Test to measure the effectiveness of a language model to engage in dialogues with a student based on the ability of the model to "speak like a teacher, understand the student, and help the student". They evaluate several BlenderBot (Roller et al., 2020) and GPT-3 (Brown et al., 2020) models on the Teacher-Student Chatroom Corpus (TSCC) (Caines et al., 2022) as well as the Educational Uptake Dataset (Demszky et al., 2021) which is a subset of the NCTE corpus that is used in this paper. Their findings found that Blender outperformed GPT-3 across all metrics, and outperformed the actual Teacher response ratings in terms of levels of uptake from student responses. However, this did not translate to outperforming teachers in terms of levels of helpfulness or the ability to create responses similar to a teacher. The BEA 2023 shared task motivated researchers to expand on this work by focusing on generating responses to student utterances from the TSCC dataset. All of the submissions were evaluated using a set of automatic dialogue evaluation metrics from (Yeh et al., 2021) and the top three models were evaluated with pairwise comparisons from human raters based on the Tack's original three categories. The best performing model, NAISTEACHER (Vasselli et al., 2023) was built on a pre-trained GPT 3.5 Turbo

(Brown et al., 2020). Their approach generates multiple teacher utterances in the form of either continuations of a previous utterance or replies to a student utterance. The responses are then re-ranked with DialogRPT (Gao et al., 2020). As part of this shared task, Hicke et al. Hicke et al. (2023) explored the use of GPT-4, as well as fine-tuning DialoGPT (Zhang et al., 2020), FlanT5 (Chung et al., 2022) and GPT-2 (Radford et al., 2019). GPT-4 outperformed their other variants. Other submissions focused on prompt engineering with open-source language models (Baladón et al., 2023) including Opt-2.7B (Gao et al., 2020) and Alpaca (Taori et al., 2023). The proposed system combines zero-shot prompts with a state-of-the-art LLM with previous research in controllable text generation and dialogue act classification to create teacher responses.

3 Methods

3.1 Overall System Design

The proposed pedagogical dialogue system consists of three major components: (1) a dialogue act segmentation and classification module (2) a dialogue policy that dictates when and how the conversational agent should intervene and (3) a response generation module that aggregates the output of components (1) and (2) into a prompt for an LLM. The dialogue system receives the most recent student utterance and the conversation history annotated with speaker labels by turn as seen in **Figure 1**. The conversation turns are fed to (1) where they are segmented into individual utterances and labeled with a dialogue act. In parallel, the most recent student utterance is fed to the dialogue policy to determine the dialogue state. The conversation state and dialogue act annotated conversation history are then aggregated into a coherent prompt for an LLM.

3.2 Dialogue Policy

A dialogue policy specifies, for each dialogue state, the actions the dialogue system can take in that state (Walker et al., 1998; Levin et al., 2000; Rieser and Lemon, 2011). Here, we leverage a dialogue policy based on an analysis of collaborative dialogue scenarios between college-aged students, where we frame the policy in terms of the dialogue acts available in each state. The students were working on a sensor-based task, in which each student becomes an expert in a particular kind of sensor - moisture, environmental, and sound, and then has to share
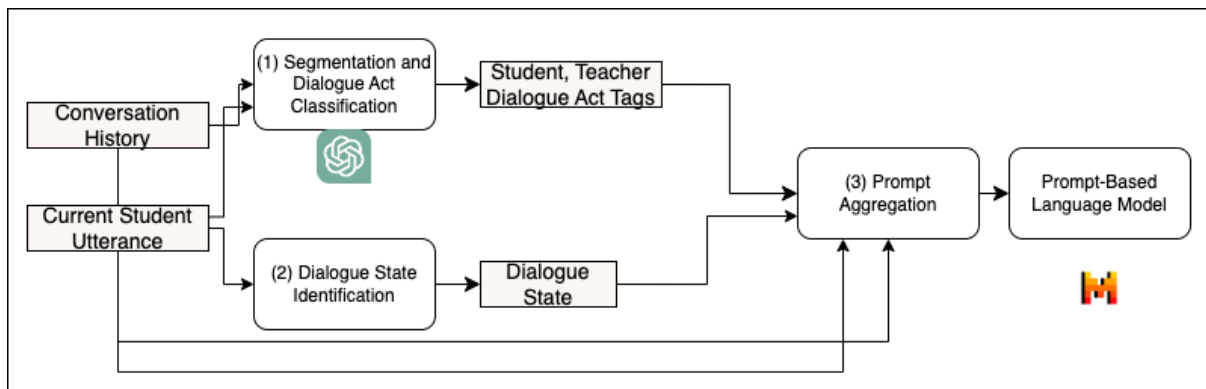
Figure 1: **Overall System Diagram**: The conversation history and current student utterance are given as input to both (1) segmentation and dialogue act classification and (2) a dialogue state identification module. They are then aggregated with the outputs of (1) and (2) into a prompt for a large language model.

the information they have learned with their group.

Education and human teamwork experts identified three broad categories of unproductive talk dialogue states during the collaboration sessions: *CONTENT*, *COLLABORATION*, and *PROCEDURAL*. The aim of our dialogue policy is to intervene in the student conversations when the conversation is in an unproductive state. Thus, each dialogue state has a set of potential dialogue actions, which are surfaced to the LLM as an additional component of the response generation prompt. The prompts are detailed in Appendix C and the complete steps for state identification are detailed in Appendix A. *PROCEDURAL* issues are identified when there is a literal String match indicating the students are confused about the next step to take to complete the task. Next, *CONTENT* issues are distinguished based on the topic feature of the NCTE dataset. The policy further divides the collaboration issues based on the number of speakers represented in the recent conversation history including *NO SPEAKER* and *SINGLE SPEAKER* categories. The NCTE dataset is segmented into only two speaker labels "student" and "teacher" without explicitly distinguishing which student is talking so we are not using these states in the experiments. For future work with multi-party conversations where speaker identification is available these are critical states to track. Additionally, since we are explicitly using non-blank utterances for the analysis, we do not have instances of the *NO SPEAKER* state. If none of the above conditions are met, the conversation is considered to be in a *FLOW* state and the student should be encouraged to continue with their reasoning.

3.3 Dialogue Act Segmentation and Classification

We use dialogue acts as a constraint for response generation as dialogue acts have been used for controllable response generation in different domains and dialogue systems (Hedayatnia et al., 2020; Ramirez et al., 2023). Following Shan et al., dialogue act labels for the NCTE corpus were obtained through few shot prompting using the gpt-3.5 turbo model. We combine the task of dialogue act classification and segmentation within the same prompt to handle longer utterances. For dialogue acts, we use the 10 ISO dialogue acts (Mezza et al., 2018b) (e.g., Inform, Set Question, Apology, etc.) and 7 Switchboard (Stolcke et al., 2000b) feedback-based dialogue acts (e.g. Backchannel, Sympathy, Appreciation) to classify and segment the NCTE dataset.

In the prompt, we provide a description of the task and instructions for the model, then for each dialogue act we give the definition and an optionally an example for certain difficult dialogue acts, and lastly we provide 19 examples of utterances that are segmented and annotated with dialogue act tags the final prompt can be found in the Appendix B. These examples came either directly from the Switchboard and ISO annotation manuals, or were handcrafted to contain the characteristics of the dialogue acts. To verify the performance of the prompt(s), we ran the prompt on a set of 100 examples, and would annotate for dialogue act and segmentation appropriateness on the teacher text and would adjust the number of examples or the definitions. Once we had the final prompt, we found that 95% of the time the model would choose an appropriate dialogue act(s). We note that the teacher dialogue acts were not always appropriate for the

conversation that could impact the performance in the downstream task.

3.4 Response Generation with Prompt Engineering

We take a modular approach to prompt engineering, wherein we dynamically construct a unique prompt for each response from component templates. These templates are injected with contextual information and desired constraints before being aggregated into a single prompt as input to the LLM. The complete set of templates can be found in **Appendix C**.

First, the baseline prompt has four components: *Preamble*, *Setting*, *Formatting*, and *Task*. We consider these to be the minimum configuration necessary to reliably produce teacher-like responses without applying any further constraints on the model's behavior. Definitions for these basic components are as follows:

- **Preamble:** This primes the model by describing the task it will ultimately be assigned after first being given the requirements and restrictions contained in the other components.

- **Setting:** This describes a persona that the model should assume, the responsibilities of its role, and environmental details, like the grade level of the students and the subject of the current lesson.

- **Formatting:** This outlines the exact format the response should take, including a soft- and hard-cap on length, expectations of attribution and quotation marks, prohibitions of explained reasoning or word count, etc.

- **Task:** This simply instructs the model to produce a response and provides the utterance to which it will respond.

Then, we add one or more of the following experimental components: *Context*, *Student DA*, *Teacher DA*, and/or *Dialogue Policy*. Unlike the baseline components, these aim to constrain model behavior. When active, they are inserted into the prompt between *Formatting* and *Task*. Definitions for these additional components are as follows:

- **Context:** This gives the model a brief conversation history, consisting of three prior utterances and who said them.

- **Student DA:** This gives the model a version of the student utterance that is segmented by the classifier (see **Section 3.3**) and the definition for each of the resulting dialogue act labels. It then instructs the model to use the given segmentation and definitions to determine what the student meant.

- **Teacher DA:** This gives the model a list of dialogue act labels (and associated definitions) found within the segmented teacher utterance, **but it does not give the actual segmented utterance**. It then instructs the model to formulate its response to match the definitions of the given labels.

- **Dialogue Policy:** This gives the model a description of the current dialogue state (see **Appendix A**) and the consequences of allowing that state to continue. It then instructs the model to formulate a response that includes the expert-recommended intervention appropriate for the given state.

Finally, we have one *implicit* constraint: the language used in the templates mimics the language used by teachers when assigning schoolwork to students. The intent is to bias the model toward seeking similar language from its training data.

4 Experiments

4.1 Dataset

For this work, we want to demonstrate the appropriateness of different prompt variations for teacher-like responses. We use the NCTE Transcripts corpus (Demszky and Hill, 2023)- a dataset consisting of 1,660 anonymized elementary school mathematics lessons. This is one of the few publicly available datasets with annotated classroom conversations. We use a subset of the dataset that is broken down into student-teacher utterance pairs where each teacher response is associated with the immediately preceding student utterance. We also provide the three prior dialogue turns for conversation history. When utterances were within the first three dialogue turns of a particular lesson, only the available turns were provided as part of the context. For evaluation, we classified the utterances and ended up with three distinct state labels present in the dataset: *CONTENT*, *FLOW*, and *COLLABORATION*. We randomly selected 100 samples consisting of 33 FLOW turns, 33 COLLABORATION,

Quality + Adherence Metrics	0.29606
Comparison to Teacher	0.24689
Best Model Response	0.21557

Table 2: Interrater Reliability based on Krippendorff's Alpha. The Quality and Adherence metrics are judged on 100 samples for all prompt variations individually. The Response Comparison metrics are based on a subset of 30 samples.

and 34 CONTENT, which is indicative of the highest subset of the datasets. All of the analysis sample utterances are then automatically segmented and classified for dialogue acts.

4.2 Prompt Variations

As detailed in **Section 3.4**, we aggregate different constraints into the prompts for the Mistral model. The goal is to discern which of these constraints leads to the most helpful, pedagogically informed outputs from the system. We start with the *Baseline* classroom response generation prompt, then provide additional constraints including: *Context*, *Dialogue State (DS)*, *DS + Student Dialogue Act (DA)*, *Student DA + Teacher DA*, and a prompt combining all the constraints (*DS, Student DA, Teacher DA*. All of the non-baseline prompts including the conversation history.

4.3 Evaluation

The generated responses are evaluated by raters on three categories: *Fluency*, *Relevance*, and *Helpfulness* using a scale from 0 to 2. Fluency to describes how natural the generated response appears to be. This is meant to be comparable to prior work that evaluates text for *Naturalness* and/or *Humanness*. *Relevance* indicates how relevant the response is to the conversation history and current student utterance, with a 1 indicating that the response is vague. Raters also annotate for *Helpfulness* to indicate whether the response helps the student or helps move the activity forward. Finally, we have a binary value to indicate how well the generated output adheres to the dialogue state-specific prompt and whether it is an acceptable based on the ISO definition of the desired output DA. We calculate Krippendorff's (Krippendorff, 2004) Alpha to gauge interrater reliability across the 600 samples evaluated by each of the three raters. The results are in **Table 2**.

5 Results Analysis

The annotation analysis is divided into three sections: output quality metrics, adherence to constraints, and teacher comparison. For the first two sections, 100 samples were annotated across all six prompt variations by three annotators. The low interrater reliability (IRR) scores in Table 1, based on Krippendorff's Alpha (Krippendorff 2004), demonstrate that this type of evaluation was difficult to come to a consensus on. This could be due to the fact that the output quality metrics were all on a score of 0 to 2, as opposed to being binary values. These three values could also be ambiguous even with the specifications provided to the annotators. To avoid this in future studies, it would be beneficial to specify a larger number of metrics that capture more fine-grained linguistic details to improve agreement across annotators. We believe the results of the study to still be valuable when considering the goal is to present an approach to designing a conversational system specifically for the educational context.

Given the low IRR values, we report the mean and standard deviation for these values to get a better sense of how the agent responses were perceived by the annotators. We see the highest agreement levels across annotators for the adherence to discourse policy. However, the adherence to dialogue act constraint did not demonstrate as high agreement. We believe this can be attributed to the fact that raters considered this to be an "acceptability" annotation - i.e. is the output aligned with an "acceptable" dialogue act for a teacher response in the given context. Additionally, when annotating the outputs were compared to the actual teacher output dialogue act tags which were annotated with our classification prompt as opposed to gold standard tags. Furthermore, the teacher utterances were often extremely lengthy making it difficult for both a model or a human to identify a single correct dialogue act.

5.1 Quality Metrics: Fluency, Relevance, Helpfulness

One of the more interesting findings was that the impact of dialogue acts on the perceived fluency of response outputs. Annotators considered responses to be less fluent (a 1 over a 2) when the model contradicted itself, provided overly formal responses, or the phrasing was considered awkward. The inclusion of Student DAs with the state led to 117%

Prompt	Fluency	Helpfulness	Relevance	DA	DS
Baseline	1.410 ± 0.596	0.917 ± 0.755	1.143 ± 0.670	0.340	0.313
+ Context	1.370 ± 0.583	1.053 ± 0.777	1.260 ± 0.626	0.490	0.360
+ DS	1.390 ± 0.564	1.070 ± 0.765	1.297 ± 0.665	0.617	0.893
+ DS, Student DA	1.613 ± 0.500	1.237 ± 0.775	1.437 ± 0.648	0.603	**0.957**
+ Student DA, Teacher DA	1.580 ± 0.563	1.287 ± 0.803	**1.530** ± 0.585	0.623	0.567
+ DS, Student DA, Teacher DA	**1.653** ± 0.503	**1.320** ± 0.798	1.450 ± 0.659	**0.670**	0.903

Table 3: Experiment results for quality metrics and adherence metrics. Fluency, Helpfulness, and Relevance are scored based on the average mean across all 100 samples and three annotators plus or minus the standard deviation. Adherence to Dialogue Act (DA) and Dialogue State (DS) constraint is based on the percentage of the time that the raters marked samples as adhering to the constraint.

Prompt	BTT	OR
Baseline	0.500	0.122
Context	**0.733**	**0.222**
DS	0.633	0.156
DS, Student DA	0.689	0.200
Student DA, Teacher DA	0.70	0.144
DS, Student DA, Teacher DA	0.678	0.156

Table 4: **BTT** represents the percentage of the time the prompt output was rated as "Better than the Teacher" response to the student utterance. **OR** represents the percentage of the samples that the prompt was rated as the best overall response of the prompt outputs.

increase in the mean Fluency score over the inclusion of the dialogue state alone. The combination of Student DAs, Teacher DAs, and Dialogue State had the highest Fluency value, suggesting that when provided more constraints, the model produced more natural responses. The prompt variation including Teacher DAs and student DAs has a much higher Fluency rating than the baseline, but lower than combining student DAs with a dialogue state suggesting the state itself contributes to higher fluency. Annotators did frequently see outputs that included dialogue act tags when the Teacher DA was included, occasionally ones that were not even mentioned in the prompt, which would need to be removed before surfacing a response in an end-to-end system. The prompt including student and teacher DAs plus state information was evaluated as having the highest mean Helpfulness score. Additionally, when the results were broken down into 2-value pairs, with [0,1] rated as "not Helpful" or 0 and 2 being scored as Helpful, this prompt was marked as helpful over 50% of the time by raters. As with the Helpfulness and Fluency ratings, we notice that the prompts that included any type of DA

information had much higher mean scores for Relevance than those without suggesting that grounding the conversation in speaker intentions leads to more relevant responses. Unsurprisingly, the baseline has the worst performance in this category given that it does not include the conversation history in the prompt. The prompt that did not include the state information was scored higher in terms of Relevance over the version with this constraint. This could be due to the model trying to attend to too many constraints at once, or this could be related to the quality of the states themselves. There are an infinite number of hypothetical states that a classroom conversation could be in that a teacher may respond to differently and in future versions of the system we intend to explore more fine-grained state-action pairs. In general, all of the response quality metrics indicate that the inclusion of DA information does lead to better output responses from an LLM.

5.2 Adherence to Dialogue Act + Policy Constraints

For annotation, raters considered an output to 'adhere' to the dialogue act constraint if it was considered an acceptable dialogue act in the context of the conversation history. The adherence rate goes up even when just including the conversation history in the prompt over the baseline suggesting that there is some implicit dialogue flow information that the model is able to learn from the history itself. However, there is a meaningful jump in performance when additional constraints are applied, including the dialogue state without any DAs. Including the DA and discourse policy information resulted in the highest rates of adherence to this constraint. However, the adherence rate is still notably less than the agent's ability to

adhere to the dialogue state constraint. In the case of our dialogue policy, the adherence rate is higher than 89% of the time when the state intervention is included in the prompt. This is also the set of annotations with the highest rates of interrater reliability. Overall, this suggests that the Mistral model was extremely good at adhering to our discourse policy when provided with the appropriate information. This is especially true in comparison to the responses when no constraints are provided, and even in the drop in performance when using the student DA and teacher DA without the discourse policy constraint. These results suggest that as we refine the discourse policy to cover a wider range of classroom situations that we will be able to output responses that will adhere to it appropriately. This finding is incredibly important when considering the need to constrain agents in the classroom to be consistent with teacher behavior and reduce the risk of providing unsafe outputs to students.

5.3 Teacher and Prompt Comparisons

In addition to evaluating each of the prompts individually, we wanted to compare them to each other, and the actual teacher responses. We selected 30 samples from the 100 annotated samples above stratified to 10 per dialogue states. The 6 model outputs were shuffled to prevent the annotators from being biased towards a particular prompt style. For these, the raters considered two questions (1) Is this response better than the teacher response? (2) What is the best model response of the 6 provided? The results in Table 5 show that all of the non-baseline prompts were considered better than the teacher response in over half of the cases selected. Unexpectedly, the condition with only the conversation history, was rated as better than the teacher most frequently and considered the best response to the student utterance most frequently. The next highest rated prompt is the combination of student DAs and the dialogue state. In further discussion of the annotations, the raters mentioned that brief responses were considered better, and that the models typically provided explicit supportive feedback such as "Keep up the good teamwork" to students more often than the teacher did. Additionally, the actual teacher responses may have been addressing earlier conversation topics or other students as opposed to the most recent student utterance. The fact that one type of constraint did not inherently improve the overall perception of the responses suggests that the system may benefit

from the use of an over-generate and rank approach in which we provide an output from all of these variations and select one to provide to the student based on a set of criteria informed by a teacher.

6 Conclusion and Future Work

The goal of these experiments was to compare the effectiveness of different constraints in the context of zero-shot prompting a language model to provide teacher-like responses to real student utterances. We evaluated the inclusion of student and teacher dialogue acts annotations as well as dialogue states in the prompts provided to Mistral. A sample of 100 utterances was selected and evaluated across 6 different prompt settings by three annotators for Fluency, Relevance, Helpfulness, and adherence to the provided constraints. The inclusion of any type of constraints showed a positive impact on all of the utterance quality metrics, but could benefit from post processing to ensure that erroneous tags are not included in the agent response to a student. We saw the highest ratings for Fluency and Helpfulness when student dialogue acts, teacher dialogue acts, and dialogue states were provided in the prompt. The prompt version with all the constraints also had the second highest value for Relevance, suggesting that more contextual dialogue information in the prompt leads to higher quality responses from the agent. When provided a dialogue state in the input, the best performing prompt adhered to the constraint in over 95% of cases. This suggests that as future research is done identifying key pedagogical dialogue states and the ability to distinguish them from one another, LLMs can be very successful in adhering to the recommended states. Furthermore, a subset of 30 utterances were compared to the teacher response from the NCTE corpus. All of the LLM prompt variations that included the conversation history were considered to be better than the actual teacher response over 50% of the time. This was largely because the annotators found the model was more likely to be directly addressing the most recent student utterance as opposed to another student, and frequently included additional supportive phrases in the response. Ultimately, these findings suggest that LLM-based conversational agents have a lot of potential for providing learners with additional support in the classroom, when provided the appropriate constraints. In future work, we aim to refine the set of states from the three present in the NCTE

128

dataset along with educators to cover a broader set of scenarios. Additionally, we would like to explore model-based approaches for identifying the dialogue states themselves. The experiments here were limited to transcript-based annotations, in future work we intend to evaluate the responses in real scenarios with students.

7 Limitations

The scope of this paper is limited by a number of factors, including the types of models used and the types of constraints evaluated. We focused on a limited set of possible dialogue states based on initial expert analysis of classroom conversations. There are more possible dialogue states that would require a different type of support in the classroom. Additionally, the annotations were based on a set of classroom transcripts. Future work should evaluate the performance of such an agent in a live learning setting.

8 Ethics Statement

The experiments that we have conducted here are intended to improve the responses generated by LLMs for the classroom setting. However, the models that we use in our experiments are trained with large datasets that may be subject to unknown biases due to the exact content of the original training materials. Our research is intended to be used as a classroom support but this assumes that teachers will not use the information collected from dialogues to assess students' grades.

Acknowledgments

References

Alexis Baladón, Ignacio Sastre, Luis Chiruzzo, and Aiala Rosá. 2023. RETUYT-InCo at BEA 2023 shared task: Tuning open-source LLMs for generating teacher responses. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 756–765, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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. *arXiv [cs.CL]*.

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.

Andrew Caines, Helen Yannakoudakis, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2022. The teacher-student chatroom corpus version 2: more lessons, new annotation, automatic detection of sequence shifts. In *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning*, pages 23–35, Louvain-la-Neuve, Belgium. LiU Electronic Press.

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. *arXiv [cs.LG]*.

Dorottya Demszky and Heather Hill. 2023. The NCTE transcripts: A dataset of elementary math classroom transcripts. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 528–538, Toronto, Canada. Association for Computational Linguistics.

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 (ACL)*.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 386–395, Online. Association for Computational Linguistics.

Umit Guz, Gokhan Tur, Dilek Hakkani-Tür, and Sébastien Cuendet. 2010. Cascaded model adaptation for dialog act segmentation and tagging. *Computer Speech & Language*, 24(2):289–306.

Behnam Hedayatnia, Karthik Gopalakrishnan, Seokhwan Kim, Yang Liu, Mihail Eric, and Dilek Hakkani-Tur. 2020. Policy-driven neural response generation for knowledge-grounded dialog systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 412–421, Dublin, Ireland. Association for Computational Linguistics.

Yann Hicke, Abhishek Masand, Wentao Guo, and Tushaar Gangavarapu. 2023. Assessing the efficacy of large language models in generating accurate teacher responses. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 745–755, Stroudsburg, PA, USA. Association for Computational Linguistics.

Juraj Juraska, Kevin Bowden, and Marilyn Walker. 2019. ViGGO: A video game corpus for data-to-text generation in open-domain conversation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 164–172, Tokyo, Japan. Association for Computational Linguistics.

Sanna Järvelä and Allyson F Hadwin. 2013. New frontiers: Regulating learning in CSCL. *Educ. Psychol.*, 48(1):25–39.

Katharina Kann, Abteen Ebrahimi, Joewie Koh, Shiran Dudy, and Alessandro Roncone. 2022. Open-domain dialogue generation: What we can do, cannot do, and should do next. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 148–165, Dublin, Ireland. Association for Computational Linguistics.

K Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Hum. Commun. Res.*, 30(3):411–433.

Mohammad Amin Kuhail, Nazik Alturki, Salwa Alramlawi, and Kholood Alhejori. 2023. Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, 28(1):973–1018.

Esther Levin, Roberto Pieraccini, and Wieland Eckert. 2000. A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on speech and audio processing*, 8(1):11–23.

Stefano Mezza, Alessandra Cervone, Evgeny Stepanov, Giuliano Tortoreto, and Giuseppe Riccardi. 2018a. ISO-standard domain-independent dialogue act tagging for conversational agents. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3539–3551, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Stefano Mezza, Alessandra Cervone, Evgeny Stepanov, Giuliano Tortoreto, and Giuseppe Riccardi. 2018b. ISO-standard domain-independent dialogue act tagging for conversational agents. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3539–3551, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Samuel L Pugh, Shree Krishna Subburaj, Arjun Ramesh Rao, Angela E B Stewart, Jessica Andrews-Todd, and Sidney K D'Mello. 2021. Say what? automatic modeling of collaborative problem solving skills from student speech in the wild. *International Educational Data Mining Society*.

Alec Radford, Jeff Wu, R Child, D Luan, Dario Amodei, and I Sutskever. 2019. Language models are unsupervised multitask learners.

Angela Ramirez, Kartik Agarwal, Juraj Juraska, Utkarsh Garg, and Marilyn Walker. 2023. Controllable generation of dialogue acts for dialogue systems via few-shot response generation and ranking. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 355–369.

Verena Rieser and Oliver Lemon. 2011. *Reinforcement learning for adaptive dialogue systems: a data-driven methodology for dialogue management and natural language generation*. Springer Science & Business Media.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot. *arXiv [cs.CL]*.

Dapeng Shan, Deliang Wang, Chenwei Zhang, Ben Kao, and Carol Chan. 2023. *Annotating Educational Dialog Act with Data Augmentation in Online One-on-One Tutoring*, pages 472–477.

Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2018. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 664–674, Melbourne, Australia. Association for Computational Linguistics.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000a. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000b. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Comput. Linguist. Assoc. Comput. Linguist.*, 26(3):339–373.

Anaã̄s 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.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Henry S. Thompson, Anne Anderson, Ellen Gurman Bard, Gwyneth Doherty-Sneddon, Alison Newlands, and Cathy Sotillo. 1993. The HCRC map task corpus: Natural dialogue for speech recognition. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Justin Vasselli, Christopher Vasselli, Adam Nohejl, and Taro Watanabe. 2023. NAISTeacher: A prompt and rerank approach to generating teacher utterances in educational dialogues. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 772–784, Toronto, Canada. Association for Computational Linguistics.

Marilyn Walker, Jeanne C Fromer, and Shrikanth Narayanan. 1998. Learning optimal dialogue strategies: A case study of a spoken dialogue agent for email. In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*.

Nick Webb and Yorick Wilks. 2005. Dialogue act classification based on intra-utterance features.

Tom Williams, Cynthia Matuszek, Kristiina Jokinen, Raj Korpan, James Pustejovsky, and Brian Scassellati. 2023. Voice in the machine: Ethical considerations for language-capable robots. *Commun. ACM*, 66(8):20–23.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Comput. Surv.*, 56(3):1–37.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Tianyu Zhao and Tatsuya Kawahara. 2017. Joint learning of dialog act segmentation and recognition in spoken dialog using neural networks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 704–712, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Tianyu Zhao and Tatsuya Kawahara. 2019. Joint dialog act segmentation and recognition in human conversations using attention to dialog context. *Computer Speech & Language*, 57:108–127.

A Dialogue State Identification

1. We classify the utterances for Collaborative Problem Solving (CPS) codes that score them on three facets: Constructing Shared Knowledge, Negotiation/Coordination, and Maintaining Team Function (Pugh et al., 2021)

2. We check if the current utterance is on-topic for the lesson based on the "ontopic" feature of the NCTE dataset

3. Check for a literal match to one of our procedural issue sentences such as "What do we do next?" then return *PROCEDURAL* issue

4. Check the utterance is ontopic and the CPS codes are above a given threshold then return *CONTENT* issue

5. Check if the CPS codes is lower than a given threshold then return *COLLABORATION* issue

6. If there are no speakers, then we consider this to be a *NO SPEAKER* collaboration issue

7. If there is a single speaker, then we consider this to be a *SINGLE SPEAKER* issue

8. If there are multiple speakers then this is a general *COLLABORATION* issue

9. If none of the above conditions are met, the conversation is considered to be in a *FLOW* state

B Dialogue Act Segmentation Prompts and Dialogue Acts

See Table 5 below.

Within the actual prompt we used 19 examples. All examples and full prompt can be found within the github repository[12].

---

[2]https://github.com/aramir62/constraints-generative-supportive-teacher-responses

The dialogue acts used from ISO are: Set Question, Propositional Question, Choice Question, Inform, Commissive, Directive, Thanking, Apology, Salutation, and Feedback (default act).

From Switchboard, the feedback dialogue acts utilized are: Signal-Not Understanding, Appreciation, Sympathy, Summarize/Reformulate, Repeat-Phrase, Acknowledge (backchannel), and Acknowledge Answer.

C Response Generation Prompts

See **Table 6** below.

D Generated Outputs

See Table 6 below.

Dialogue Act Segmentation Prompt
You are given an utterance to label a dialogue act with, and certain utterances need to be segmented if needed. You'll segment the utterance into as many chunks needed to capture all the dialogue acts, but if there are two chunks in a row with the same dialogue act combine them. Use punctuation and clause separators as a way to consider if this is a new topic or idea that would be labeled with a different dialogue act. Choose only one dialogue act per segment and you have to choose one. Definitions and examples of each given below. Only choose dialogue acts from this set, dialogue act names are contained in <>.   <Set Question>: A question that focuses on the speaker wanting to know certain information often contains "wh" at the beginning such as "Where are you going?", "What did Kevin do yesterday?"   <Propositional Question>: A question where the speaker wants to know if something is true or false, such as "Do you know what time it is?"   <Choice Question>: A question that provides a list of options for someone to choose from, usually contains or and requires the other speaker to choose something, such as "Do you like peanut butter or chocolate more?"   <Inform>: A statement that is providing information. Described as informative information to provide context, knowledge, and information about a topic.   <Commissive>: The speaker will express a commitment to performing an action, such as "I will go home at 1pm"   <Directive>: The speaker is directing or suggesting what will need to be committed to, such as "You need to go left then right down the hall"   <Thanking>: A sentence that is expressing gratitude, such as "Thanks"   <Apology>: An expression acknowledging regret or remorse towards an individual's actions, such as "I'm sorry"   <Salutation>: An utterance made as a greeting or acknowledgment of another's arrival or departure, such as "Hello"   <Signal-Not Understanding>: The speaker is expressing that they did not understand what was said prior. Or is using some form of non verbal language.   <Appreciation>: The speaker is expressing feedback by providing appreciation towards the other speaker.   <Sympathy>: The speaker is sympathetic towards the other speaker and is expressing this within the utterance.   <Summarize/Reformulate>: The speaker summarizes or reformulates what was said before to demonstrate understanding of what was heard.   <Repeat-phrase>: The speaker repeats back what was said beforehand.   <Acknowledge (backchannel)>: The speaker expressing acknowledgement by using backchannels such as "Mmm hmm", "Mm yeah", "Uh huh"   <Acknowledge Answer>: The speaker provides acknowledgement of what was said or asked prior by providing an answer.   <Feedback:> An utterance that provides or elicit information about the type of understanding and processing of what was said prior, use this as a default if the other sets do not match and put as the lowest priority.   Utterance: You need to give me your ideas and then I need to see whether that would sell in the marketplace.   Output: You need to give me your ideas<Directive> and then I need to see whether that would sell in the marketplace<Commissive>   Utterance: {utterance}   Output:

Table 5: **Dialogue Act Segmentation Prompt** with 1 example (prompt used for labeling used 19 examples.)

Preamble:	
You will be given the task of generating a realistic response to a given statement or question. In order to complete this task successfully, you must pay careful attention to the following requirements and limitations. Failure to do so will result in failure of the task.	
Setting:	
For context, the given statement or question will have been spoken by a **{LEVEL}** student in the setting of a **{SUBJECT}** classroom. The student is working on an in-class assignment with a group of their peers. You will generate a response to what the student said as if you are the teacher of this **{SUBJECT}** class. As a teacher, your response must be: (1) kind, (2) helpful, (3) encouraging of collaboration, (4) relevant to the subject of **{SUBJECT}**, (5) factual, and (6) befitting the role of a professional classroom teacher. Please limit your response to the content of the student statement or question. These guidelines may help you:	
Formatting:	
Please format your response to the given statement or question as a SINGLE direct quote from the **{SUBJECT}** teacher whose role you are playing, including quotation marks. Please try to limit your response to **{SOFT_CAP}** words or less words if you can. However, if you are unable to accurately and coherently respond to the student's statement or question in **{SOFT_CAP}** words or less, you may use up to **{HARD_CAP}** words. Any response longer than **{HARD_CAP}** words will be considered a failure for the purpose of this task. Please do NOT include a word count in your response. Please do NOT provide any additional reasoning, explanations, or context along with your response. Please preface your response with the text "Response:" OUTSIDE of the quotation marks, but do NOT include any additional or alternative speaker attributions. Please do NOT preface or follow your response with ANY text other than the quotation marks and attribution required above.	
Student DA & Teacher DA:	
The following are definitions for dialogue act labels, which are useful for understanding text: **{LABEL_DEFINITIONS}** To help you understand the student, their statement or question has been annotated with dialogue act labels. These labels have been inserted AFTER the section of text they describe, and each label will correspond to one of the definitions shown above. What follows is the annotated student quote: **{STUDENT_SEGMENTATION}** Furthermore, you should formulate your response such that it matches the definition(s) for the following label(s): **{TEACHER_LABELS}** However, the dialogue act labels are exclusively for your own use in understanding what the student has said and in formulating your response. Please do NOT include ANY actual dialogue act labels in your response.	
Dialogue Policy:	
{STATE_DESCRIPTION}[3] **{CONSEQUENCES}** **{INTERVENTIONS}**	
Context:	
To give you some insight into the on-going discussion, the following conversation history is provided to you: **{HISTORY}**	
Task:	
The following student quote is what you will be responding to: **{STUDENT_UTTERANCE}** Please generate the teacher's response according to the requirements and limitations provided above.	

Table 6: **Response Generation Prompt** with all 8 component templates.

State	Action/Prompt
Content Issue	This group of students appears to be struggling with some aspect of the lesson material. The purpose of this activity is for them to learn the material; however, if they are struggling to understand the new information or have forgotten prior learning, they may be unable to finish the assignment. In order to help them make progress toward their current objective, your response should politely encourage them to compare notes and make sure that they agree on the lesson material.
Collaboration Issue	This group of students appears to be actively speaking, but none of them are making meaningful progress toward their current objective. This could be a problem, because class time is limited. If they spend too much time stuck or off-task, they may not complete their group assignment by the end of the class period. Your response should gently remind them of their current objective, politely encourage them to work collaboratively to achieve it, and perhaps suggest a way for them to get back on task
Procedural Issue	The students are struggling to understand the steps required to complete the task. Please encourage them to refer to the Sensor Immersion tutorials. This group of students appears to have found a good working rhythm and are making meaningful progress toward their current objective. They are doing a good job of collaborating with one another, and you do not need to intervene. Therefore, your response should praise their enthusiasm and encourage them to remain on-task and continue to make steady progress.

Table 7: A breakdown of the action prompts used for the content, collaboration, and procedural issue dialogue states. The procedural issue is not present in the NCTE dataset and has reference to the Sensor Immersion task used in the dialogue policy creation sessions.

Student Utterance [Flow]	Conversation History
So, not in the envelope?<Feedback>	teacher: "Anybody? Nobody? What did you use them for, Student C."
	student: "To help us with our fractions."
	teacher: "To help you with your fractions. Do you guys understand that sometimes it's easier to just see it and touch it and hold it? Today, the job that we have is that we are going to learn to simplify fractions and we are going to use these pieces to help us to simplify them. So, by the time I am done, I have 1, 2, 3, 4, 5, 6, 7, 8, 9, piles of fraction pieces. And one of them is one whole strip. And all of your scraps need to go into the blue recycling bin, please, which I just put right here. I'm not so worried about your names going on the backs of them, boys and girls, because you have your own little toolbox, the envelope you are going to keep them in, so your name does need to go on the envelope."

Prompt Variation	Generated Response
Actual Teacher Response	I'm sorry, what?<Signal-Not Understanding>/ It does go on the envelope.<Inform>/ We need to get moving here.<Directive>/ It's the noise.<Inform>/ We don't need to do that.<Directive>/
Baseline	"That's correct, the problem does not involve finding the area or volume of an envelope."
+ Context	"That's correct. The fraction pieces themselves don't need names on them, but be sure to label your envelopes."
+ DS	"That's correct, Student. The names don't need to be on the fraction pieces themselves, only on the envelopes. Keep up the good work."
+ DS, Student DA	"Great question! Yes, you're correct, any scraps or unused pieces don't go in the envelope. Keep up the good work, team!"
+ Student DA, Teacher DA	You're correct, those pieces that don't fit in the envelope should be placed in the recycling bin. Keep up the good work, team!
+ DS, Student DA, Teacher DA	"Yes, that's correct. The scraps don't need to go in the envelope, they can go in the blue bin." <Inform, Directive>

Student Utterance [Collaboration Issue]	Conversation History
That's not how I write my fractions though so<Feedback>/	teacher: "Okay, all right. And then if you finish, just ask yourselves. Come up with a fraction and ask yourself, you know, how you can compare." teacher: "Okay, hard to write on those little lines, isn't it?" teacher: "Bless you."

Prompt Variation	Generated Response
Actual Teacher Response	All right,<Acknowledgement>/ are you girls finished?<Set Question>/
Baseline	"I understand that everyone has their own way of representing fractions. Let's discuss how we can all use the same convention for this assignment to ensure clarity."
+ Context	"I understand, everyone has their own way of writing fractions. Let's compare our methods and learn from each other."
+ DS	"Understood. Let's try using the standard fraction format for consistency within your group to ensure effective comparisons."
+ DS, Student DA	"I understand that everyone has different ways of representing fractions. Let's collaborate and share our methods to help us all learn. Let's focus on understanding the concept of comparing fractions."
+ Student DA, Teacher DA	"It's perfectly fine if we each have slightly different ways of writing fractions as long as the meaning is clear. Let's discuss how we can compare them effectively."
+ DS, Student DA, Teacher DA	"It's great that we all have different ways of expressing ideas. Let's see if we can use each other's methods to help us understand how to compare fractions effectively. Let's continue working together."

Student Utterance [Content Issue]	Conversation History
Because a straight angle goes on, on both sides.<Inform>/	teacher: "Let's see. All right. Good. Straight angles. Let's get our pencils ready because this is one you need to do. Some of you may have heard of this. How many of you all have heard of a straight angle? Very small. All right. A straight angle is a line that goes infinitely in both directions. It measures 180 degrees. Okay. Yes, Student L? All right. So again, which is not a straight angle? Any others? What do you think, based on definition?" student: "Number two." teacher: "How do you know number two is not a straight angle?"

Prompt Variation	Generated Response
Actual Teacher Response	It goes on forever. There's a line, and it goes on forever on both sides, absolutely.<Inform>/ Of course, it's a ray so it only goes in one direction.<Inform>/ Now we can talk about our main guy today.<Inform>/ Our main guy today is a triangle.<Inform>/ We all know triangles, right?<Acknowledge Answer>/
Baseline	"That's an interesting way to think about it. Can you clarify what you mean by 'a straight angle goes on, on both sides' in the context of geometry?"
+ Context	"That's correct, Student. A straight angle does indeed go on indefinitely in both directions. Well done for applying the definition."
+ DS	"That's correct, Student, a straight angle does indeed extend indefinitely in both directions. Let's discuss with your group to ensure everyone agrees on this definition."
+ DS, Student DA	"That's correct, a straight angle extends infinitely in both directions. Let's compare our definitions and work together to find other angles in the diagram that are not straight angles."
+ Student DA, Teacher DA	"That's correct, Student. A straight angle extends indefinitely, meaning it continues on both sides, which is why it measures 180 degrees."
+ DS, Student DA, Teacher DA	"That's correct. A straight angle continues indefinitely on both sides, making it an important concept in geometry. Let's discuss this with your group and ensure everyone agrees on this definition."

Table 8: A comparison of the generated utterances from the different prompt variations for each of the dialogue states.