NAISTeacher: A Prompt and Rerank Approach to Generating Teacher Utterances in Educational Dialogues

Justin Vasselli[†,††] **Christopher Vasselli**[††] **Adam Nohejl**[†] **Taro Watanabe**[†]
[†] Nara Institute of Science and Technology [††] Serpenti Sei
[†] {vasselli.justin_ray.vk4, nohejl.adam.mt3, taro}@is.naist.jp
[††] chris@serpentisei.com

Abstract

This paper presents our approach to the BEA 2023 shared task of generating teacher responses in educational dialogues, using the Teacher-Student Chatroom Corpus. Our system prompts `GPT-3.5-turbo` to generate initial suggestions, which are then subjected to reranking. We explore multiple strategies for candidate generation, including prompting for multiple candidates and employing iterative few-shot prompts with negative examples. We aggregate all candidate responses and rerank them based on DialogRPT scores. To handle consecutive turns in the dialogue data, we divide the task of generating teacher utterances into two components: teacher replies to the student and teacher continuations of previously sent messages. Through our proposed methodology, our system achieved the top score on both automated metrics and human evaluation, surpassing the reference human teachers on the latter.

1 Introduction

The shared task for BEA2023 was to generate teacher utterances in an educational dialogue, specifically one between an English language learner and their language teacher (Tack et al., 2023).

The data was collected from one-on-one English lessons between real teachers and students conducted over a chat application. The data for the task consists of fragments of these dialogues, with the goal of predicting the next teacher utterance.

Inspired by a commonly used practice in machine translation (Och and Ney, 2002; Shen et al., 2004; Lee et al., 2021), our system generates multiple candidates and reranks them. Given the high level of fluency required for this task, we began with a pretrained language model (`GPT-3.5-turbo`) rather than training one from scratch.

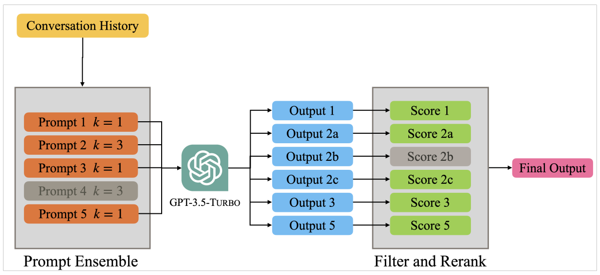


Figure 1: The NAISTeacher system overview. Prompts are chosen for the ensemble based on the role of the speaker of the final utterance of the conversation history. $k = 1$ and $k = 3$ refer to the number of candidates solicited by the prompt (one and three respectively).

An overview of the system is outlined in Figure 1. First the prompts for the prompt ensemble are chosen based on the conversation history. The prompt ensemble is sent to `GPT-3.5-turbo` to generate a set of candidate responses. During a post-processing step some responses are flagged as inappropriate for referring to pronunciation, responding as a student, or containing profanity. These inappropriate responses are filtered out before reranking. The response chosen by the reranker is returned.

Our prompt ensemble consists of a mixture both zero and few shot prompts. For the few shot prompts, we experimented with several different ways of selecting the examples to provide, but the most effective was an iterative approach that, to our knowledge, is novel. This approach works in two steps: The first step is to generate the candidate teacher utterances for all conversations using the zero-shot prompt and score them. In the second step, the highest and lowest scoring responses are integrated into the prompt as positive and negative example responses.

To generate multiple candidates, we instructed the model to return multiple possible responses with a single prompt similar to Aumiller and Gertz (2022). We found this to be a very effective technique because it resulted in a diverse set of candidates to select from in the reranking stage.

We tested our methods on a held-out section

Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 772–784
July 13, 2023 ©2023 Association for Computational Linguistics

of the training data which contained reference responses using BERTScore (Zhang et al., 2019) and DialogRPT (Gao et al., 2020). We found the iterative few-shot approach increased the overall score of the data on both automated metrics. However, manual inspection of the output showed that this approach works better for some cases than others.

Teacher response or teacher continuation?

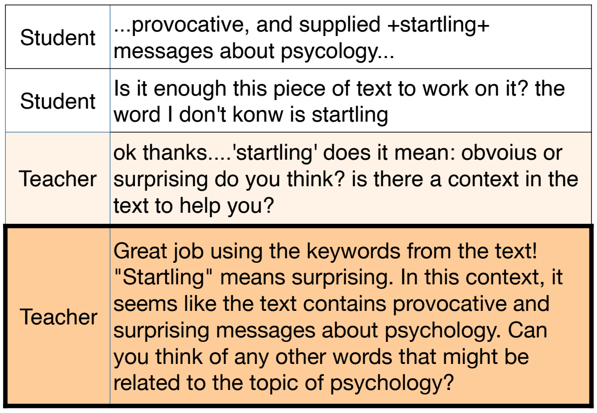
As the data was collected from real-time chats between real people, the turns did not alternate consistently as they do in most dialogue systems. The task of generating a teacher continuation – that is, a teacher follow-up message when that same teacher was the last to have taken a turn in the conversational history – was particularly challenging. This phenomenon occurs in 38% of the training data, and 40% of the development data. As systems are trained to alternate turn taking, this led to many of the generated responses erroneously sounding like they were either uttered by the student, or showed a lack of awareness of the immediately preceding utterance.

Figure 2 shows one such conversation and an example of a problematic generated response. The teacher first asks the student to look at the text and think about if the word “startling” means obvious or surprising. The AI generated follow-up gives the answer away immediately (`"Startling" means surprising`), which overrides the original goal: to encourage the student to find the answer themselves.

To address this, we split the data into two tasks based on who spoke last: **teacher reply** for when the context ended with the student, and **teacher continuation** for when the context ended with the teacher. We evaluated each task separately during development to design a set of prompts for each.

2 Data

Although we explored augmenting with data from other sources, in the final system we only used the data made available for this shared task, the Teacher-Student Chatroom Corpus (Caines et al., 2020, 2022). The training data consists of 2747 partial conversations between a student learning English and their teacher. We split out a random sample of 275 conversations (10%) to use as validation data. Our experiments were carried out on this split because the development data supplied by the shared task did not contain reference answers.



Student	...provocative, and supplied +startling+ messages about psycology...
Student	Is it enough this piece of text to work on it? the word I don't konw is startling
Teacher	ok thanks....'startling' does it mean: obvoius or surprising do you think? is there a context in the text to help you?
Teacher	Great job using the keywords from the text! "Startling" means surprising. In this context, it seems like the text contains provocative and surprising messages about psychology. Can you think of any other words that might be related to the topic of psychology?

Figure 2: An example of a teacher continuation with the response to the zero-shot (1) prompt.

Besides a single run of our system on this data to submit to CodaLab, we did not use the 305 conversations in the supplied development dataset. The test data contained 273 conversations.

3 System

Our system generates a number of responses by prompting `GPT-3.5-turbo` using the default parameters, and chooses among them using a DialogRPT-based reranker. NAISTeacher uses several different prompts, ranging from general to those targeting specific scenarios.

3.1 Zero-shot prompt

The base zero-shot prompt gives `GPT-3.5-turbo` the conversation history, along with scaffolding to encourage it to answer in a teacher-like way. The prompt reads as follows:

(1) The following is a partial conversation between an English language learner and their teacher:

> *(conversation)*
>
> Can you give an example teacher follow-up to their previous message that would be helpful for the language learner? The message should be concise, and worded simply. It should either encourage the continuation of the current topic or gracefully transition to a new teacher-provided topic. Questions should be specific and not open-ended. Try to not sound like an assistant, but a teacher, in charge of the flow of the lesson.

The prompt went through several iterations to address problems such as answering questions too directly and sounding too much like an assistant. It started as a simple one-sentence prompt and we manually tested additions one-by-one on a subset of 3–5 conversations to address a number of challenges observed in the responses:

– the response being too thorough or technical;
– the end of the response not engaging the learner;
– questions being too open-ended, i.e. `Is there anything else you're unsure about or need help with?`

Our final prompt requests responses that are concise, encourage student engagement, and sound like a teacher rather than an assistant. We found that responses to this prompt tended to be a manageable length and level of detail, and they invite the student to engage with the conversation further and think about the topic in more depth.

See Appendix A for details on the iterations we tested for the zero-shot prompt.

3.2 Few-shot prompts

For the few-shot prompts, we provide one example teacher response that emulates the type of response we want, and one example of a response we would like to avoid. We used three different methods for choosing the positive and negative examples: handcrafted, generative, and iterative. See Table 1 for the results of the three different methods.

3.2.1 Handcrafted examples

We experimented with short manually written examples such as the following:

(2) *Concatenation of prompt (1) and the following:*

Good example: 'Can you make a sentence using 'within'?' Bad example: 'Do you have any questions about prepositions?'

The goal of these examples was to get the system to ask questions to maximize learning for the student, but to not allow those questions to get so general that the student is most likely to respond with a short, one-word answer.

3.2.2 Generated examples

Inspired by chain-of-thought prompting (Wei et al., 2022), we asked `GPT-3.5-turbo` to first consider what makes a good teacher response, and then integrated the answer back into the prompt.

For the case of teacher reply, we used `GPT-3.5-turbo` to generate a prompt with a 1-shot example pair, one positive, one negative. First `GPT-3.5-turbo` was asked:

(3) How does a teacher sound when responding to a student? What kinds of things would teachers say that chatbots would not? What do they not say? In your response provide an example of a response that sounds like a teacher and one that sounds like a chatbot? Respond succinctly.

	BERTScore	DialogRPT
(1) Zero-shot, $k = 1$	70.91	36.51
(1) Zero-shot, $k = 3$	70.04	36.27
(2) Handcrafted 1-shot, $k = 3$	70.31	35.46
(4) Generated 1-shot, $k = 1$	70.88	38.26
(5) Iterative 1-shot, $k = 1$	**71.55**	**40.94**
Reference	1.0	32.81

Table 1: The results on the full test set of candidates generated with several different prompts. k is the number of candidates solicited by the prompt. For $k > 1$ the score is the average of all the candidates with no reranking.

The answer to this question was integrated into the zero-shot prompt. Here is the final prompt with the GPT generated portion in bold.

(4) The following is a partial conversation between an English language learner and their teacher:

(conversation)

They are in the middle of a lesson. Can you give a possible way the teacher could respond?

Remember: **A teacher typically sounds knowledgeable, authoritative, and focused on guiding and instructing students. They may use formal language and provide detailed explanations. Teachers often offer constructive feedback, encourage critical thinking, and ask probing questions to stimulate learning.**

Example of a teacher-like response: "That's a great observation, but let's delve deeper into the topic. Can you provide some evidence to support your claim?"

A chatbot, on the other hand, may sound more informal and conversational. It tends to provide general information or brief responses without much elaboration.

Example of a chatbot-like response: "Interesting! Tell me more."

Teachers typically avoid expressing personal opinions or biases. They also refrain from engaging in casual banter or unrelated conversations to maintain a professional and educational atmosphere.

3.2.3 Iterative examples

The third method of generating examples for the prompt was an iterative approach. For this, we first used the zero-shot prompt (1) to generate responses for all conversations in the data. Next we used DialogRPT to score the responses and selected the highest and lowest scoring responses as the positive and negative examples respectively. In the final prompt we do not provide the full conversation context that led to the example, rather we use just the positive and negative response examples themselves appended to the end of the zero-shot prompt. The final prompt is as follows:

(5) *Concatenation of prompt (1) and the following:*

> Here is an example of an exceptional teacher follow-up:
> "Great job, student! Just a small correction, we should use the present tense verb "built" instead of "build" since the construction has already been completed. So the correct sentence is: "The International Space Station is built by NASA." Keep up the good work! Now, let's move on to a new topic - let's talk about your favorite hobbies. Can you tell me what activities you enjoy doing in your free time?"
>
> Here is an example of a poor teacher follow-up:
> "That's an interesting observation about poshness. Can you think of any examples of British accents that might be associated with poshness?"

The idea behind this was to optimize for high scoring prompts on DialogRPT, and the results show an improvement in average DialogRPT scores. See Table 1 for the automatic evaluation of the responses from each of these prompts.

3.3 Prompting for multiple candidates

By modifying `Can you give an example` ($k = 1$) to `Can you give three examples` ($k = 3$), we were able to illicit three replies at once. While originally implemented in an attempt to save both time and money when generating multiple candidates, this technique had multiple unexpected positive effects: it produced shorter responses in line with the length of the reference sentences, increased the diversity of the output as compared to running the same prompt twice, and allowed us to filter out candidates with profanity or references to things that would be inappropriate in a text chat (i.e. pronunciation practice) in a post-processing step (see 3.5 Post-processing).

The real teachers responded with comparatively short responses: the mean response length of the references was 23 words. Without specifying length requirements, `GPT-3.5-turbo` would return longer, more thorough responses, averaging over 35 words. By requesting three responses at once, the options shortened naturally to just over 23 words in the zero-shot case. While shorter responses may not work as well for other tasks, on this task, the shorter responses more closely matched the length of the reference sentences. See Table 2 for a comparison of sentence length and automatic evaluation.

	avg. characters	avg. words
(1) Zero-shot, $k = 1$	205.16	35.55
(1) Zero-shot, $k = 3$	137.09	23.67
Reference	126.26	23.01

Table 2: The average length of the responses of the zero-shot prompt with $k = 1$ and $k = 3$ compared to the reference teacher responses.

	distinct-1	distinct-2
(1) Zero-shot, $k = 1$	0.62	0.84
(1) Zero-shot, $k = 3$	0.77	0.91

Table 3: The average percentage of distinct unigrams (distinct-1) and bigrams (distinct-2) present in the candidate sentences.

In addition to making the responses more concise, requesting multiple candidate responses at once also introduced more diversity to the responses compared to requesting a single response three times. There are many possible ways for a teacher to respond in a given situation, and generating many candidate responses allows the system to choose the best one. However, there is not much value to be gained from choosing between very similar candidates. When the candidates are more diverse, there is more chance of generating a really high quality response. We found that the responses generated by the $k = 3$ prompts were more diverse compared to multiple inferences using the same prompt.

To measure diversity, we calculated the distinct unigrams and bigrams present in the candidates, normalized by dividing by the total number of words in the candidates following Li et al. (2015). The results shown in Table 3 demonstrate that there is very little overlap between the candidates generated by the $k = 3$ prompts.

While using $k = 3$ prompts generates shorter and more diverse candidate, it does come at the cost of a slight performance hit on the automated metrics. For further investigation into possible causes of this see 5.2 DialogRPT Length Bias Investigation. We used a combination of $k = 1$ and $k = 3$ prompts to balance the output of the full system.

3.4 Adaptations for teacher replies vs. continuations

Upon manual evaluation of the output, we found that some prompts, including the iterative 1-shot prompts (5), are better suited to generating teacher replies (the teacher turn following a student turn) than teacher continuations (the teacher turn following a teacher turn). We split the task of responding

	BERTScore	DialogRPT
(1) Zero-shot, $k = 1$	71.34	37.25
(1) Zero-shot, $k = 3$	70.32	36.21
(2) Handcrafted 1-shot, $k = 3$	70.62	35.53
(4) Generated 1-shot, $k = 1$	71.52	39.25
(4) Generated 1-shot, $k = 3$	70.98	37.39
(5) Iterative 1-shot, $k = 1$	**72.15**	**41.53**
(6) Reply, $k = 3$	70.82	34.67
(13) Reply long, $k = 3$	70.82	34.68
(7) Targeted transition, $k = 3$	70.32	37.21
Reference	1.00	32.31

Table 4: The results on the subset of the test data where the last speaker was the student (teacher reply).

to the student from the task of generating teacher follow-up utterances and evaluated the general systems separately on this subset of the data. At inference time, the system selects which prompts will be used for the ensemble based on the speaker in the final turn of the provided conversation history; if the role was a student, it chooses the prompts for teacher reply, and if the role was a teacher, it chooses the prompts for teacher continuation.

3.4.1 Teacher reply

The case of teacher reply can be thought of as the default case. The prompts described thus far generated reasonable responses for this case. Table 4 contains the results of the different prompts used for teacher reply evaluated only on the subset of the test data where the generated teacher utterance is responding directly to the student. The iterative 1-shot (5) prompt scored highest on both automated metrics, but many of the prompts generated several high quality candidates that were chosen for the final output of the system.

Two additional zero-shot prompts were engineered specifically to target the case of teacher reply, and were only used for conversations where the final utterance was from the student:

(6) Here is a partial conversation between a student and their teacher during a private English lesson:

(conversation)

Can you give three possible ways the teacher could respond to continue the lesson? Use Simple English. While the conversation might be about culture or other topics, the point is to practice English
Each teacher response should:
1. Acknowledge what the student said, and demonstrate understanding.
2. Be helpful to the student, without answering directly. Give hints to help the student think for themselves.
3. Encourage the student to respond with an exercise or question.
Respond without preamble, just number them.

This prompt was engineered to target two of the three criteria for human evaluation from Tack and Piech (2022): Does it sound like it understands the student? Is it helpful for the student?

The second prompt was engineered to allow for changes of topic in the conversation. Originally this prompt was applied to the full dataset. However, we found that the responses were less effective in the case of teacher continuation, and so in the end they were used only for the case of teacher reply.

(7) The following is a partial live chat between a teacher and a student learning English. They are in the middle of a lesson. Can you provide 3 possible ways the teacher could wrap up the current conversation and start an exercise or new topic of discussion?

Remember: Teachers often use specific language and techniques that chat bots have difficulty replicating. For example, they may ask open-ended questions to encourage critical thinking and engagement, provide specific feedback on a student's work, or offer personalized guidance based on a student's strengths and weaknesses. Teachers have a specific agenda for each lesson, such as practicing a specific grammar point or vocabulary. Try to understand what's happening in the conversation and what the teacher's goal is for the lesson.

If the goal is unclear, you can assume that the teacher wants to move on to a new topic or exercise.

Don't ask questions that are overly general such as "Is there anything else you'd like to talk about?"

Here is the conversation so far:
(conversation)

3.4.2 Teacher continuations

The task of generating teacher continuations proved more challenging than generating teacher replies. As dialogue systems are not typically trained on this task, it is particularly prone to producing spurious student responses between turns, i.e. responding as the student before providing a teacher response. One of the most common problems that arose in this subtask was that the generated candidate continuations would try to respond to the previous utterance as if it were a different speaker. To address this, several zero-shot prompts were engineered to cover possible reasons for a teacher to send a follow-up message before the student takes a turn. These prompts were carefully crafted during the development phase to ensure there were fewer spurious student responses.

The results of the generally applied prompts were manually evaluated to judge the appropri-

	BERTScore	DialogRPT
(1) Zero-shot, $k = 1$	69.89	34.82
(1) Zero-shot, $k = 3$	69.47	36.38
(2) Handcrafted 1-shot, $k = 3$	69.69	35.31
(8) Continue, $k = 3$	70.09	36.29
(14) Continue long, $k = 3$	69.49	34.76
(10) Generated 1-shot, $k = 3$	69.01	32.82
(11) Exercise, $k = 1$	**70.20**	**36.42**
(12) Conversation, $k = 1$	69.37	32.11
Reference	1.00	33.81

Table 5: The results on the subset of the test data where the last speaker was the teacher (teacher continuation).

ateness of the output for this subtask. Despite a high performance on the automated metrics, we removed the iterative 1-shot prompt (5) responses from consideration in the teacher continuation case. This decision was made because the candidates frequently sounded as if the student said something between the teacher utterances. The results of evaluation on the teacher continuation subset of the test data are shown in Table 5[1].

The first prompt generated specifically for use in teacher continuations was a simple one.

(8) Here is a partial conversation between a student and their teacher during a private English lesson:

(conversation)

Can you give three possible ways the teacher could continue their response? Use simple English.

Similar to the prompt generated for the teacher reply (4), we used `GPT-3.5-turbo` to generate a detailed prompt for the case of teacher continuation. First `GPT-3.5-turbo` was asked:

(9) In the following conversation, the teacher has already sent a message. As this is a live chat, they want to send another message right away, before the student has a chance to reply. What might be some reasons why they want to follow-up on their previous message?

The answer was embedded in a new zero-shot prompt:

(10) The following is a partial live chat between a teacher and a student learning English. They are in the middle of a lesson, and the teacher has already sent a message, but wants to follow-up. **There could be various reasons why the teacher wants to follow-up on their previous message before the student has a chance to reply. Here are some possibilities:**

1. The teacher may have realized that their previous message contained some inaccuracies or omissions, and they want to correct or clarify their statement to avoid confusion.
2. The teacher may have received new information or thought of a better way to explain something, and they want to add to their previous message to provide a more complete answer.
3. The teacher may want to check if the student has any further questions or needs more explanation on the topic, and they want to encourage further discussion by sending a follow-up message.

Regardless of the reason, the teacher's follow-up message can help ensure that the student fully understands the topic being discussed and feels comfortable asking questions and engaging in the conversation.

Can you provide 3 possible follow-up messages the teacher could write?

Use simple English. The response should sound like a teacher, not an assistant. Good example: ’Can you make a sentence using ’within’?’ Bad example: ’Do you have any questions about prepositions?’. The response should be helpful for the student and show that the teacher understood the student.
Here is the conversation so far:

(conversation)

3.4.3 Specific teacher continuation scenarios

Two prompts were designed for teacher continuation. The first prompt is used when the teacher has not provided an exercise or question for the student to respond to. In a lesson, it is typically the teacher's responsibility to keep the student engaged, the conversation flowing, and the lesson on track. With this in mind, we asked `GPT-3.5-turbo` to check if the teacher has already asked a question, and if not, to provide one.

(11) Here is a partial conversation between an English student and their teacher:

(conversation)

In the last utterance, did the teacher ask a question? If not, please provide one that would be appropriate. If they were in the middle of an exercise, what should they say to continue the exercise? The question or prompt should be simple. Don't be too verbose or open ended. Good example: "What else is ’surprising’?" Bad example: "Is there anything else you'd like to know?"

Respond in the following format:
Teacher asked a question: (yes/no)
Question or prompt:

Similarly, a prompt was generated for the case that the teacher and student were engaged in more casual chitchat rather than exercises.

(12) Here is a partial conversation between an English student and their teacher:

---

[1] While we report them here, we consider the DialogRPT scores to be unreliable as the models were not trained to evaluate this subtask.

(conversation)

Were they in the middle of a conversation? If so, what should the teacher say to continue the conversation? The question or prompt should be simple and not use terminology such as 'collocations'. Don't be too verbose or open ended. Good example: "What else is 'surprising'?" Bad example: "Is there anything else you'd like to know?"

Respond in the following format:
Conversation: yes or no
Teacher:

Both of the above prompts required a bit more post-processing, but explicitly requesting a format in the prompt simplified this task.

3.5 Post-processing

The raw outputs of `GPT-3.5-turbo` contained inconsistent formatting, sometimes including quotes around the sample response, or sometimes prefixing `Teacher:` to the reply. The prompts that asked for multiple responses resulted in a numbered list, sometimes formatted `1:`, `2:`, and sometimes `1)`, `2)`. Occasionally this would include preamble such as `The teacher could reply:`. Prompts (11) and (12) both specified a pattern of output, and required slightly different post-processing to extract the relevant information and text. In the case of (11), if an exercise was already provided then there was no need to save the suggested candidate.

Post-processing was done on all of the `GPT-3.5-turbo` outputs to make the format more consistent and to separate the replies when multiple were requested. Separating the replies on the $k = 3$ prompts was as simple as splitting on the new line, discarding lines that did not start with a number, and removing the numbers with the regular expression `/^\d+[\.)]\s+/`. If the remaining text was enclosed in quotes, the quotes were removed. If the remaining text started with `teacher:`, the prefix was removed. If the string started with `student:`, the entire candidate response was flagged as a student utterance and removed. If the response included any of the following phrases that indicate a request for a verbal response, we removed it from the list of candidates:

```
try repeating
repeat after me
practice pronunciation
```

	contributions to final
(1) Zero-shot, $k = 1$	20
(2) Handcrafted 1-shot, $k = 3$	35
(4) Generated 1-shot, $k = 1$	22
(5) Iterative 1-shot, $k = 1$	72
(6) Targeted reply, $k = 3$	12
(7) Targeted transition, $k = 3$	40
(8) Continue, $k = 3$	17
(14) Continue long, $k = 3$	14
(10) Generated 1-shot, $k = 3$	27
(11) Exercise, $k = 1$	11
(12) Conversation, $k = 1$	3
Total	273

Table 6: The number of responses from each prompt that were chosen by the reranker for the final output.

For the final system, which chooses between utterances generated by several different prompts, each candidate response was run through a profanity filter[2] and discarded in the case of a profanity being detected.

3.6 Reranking

We used a very simple reranker that chose the candidate response with the highest DialogRPT score. The final score was calculated as a composite of subscores.

$$\begin{aligned} D_{\text{final}} = {} & (D_{\text{updown}} + 0.48 D_{\text{depth}} - 0.5 D_{\text{width}}) \\ & \times (0.5 D_{\text{vs-random}} + 0.5 D_{\text{vs-machine}}) \end{aligned} \tag{1}$$

Each of the scores was calculated with a different HuggingFace model[3]:

- microsoft/DialogRPT-updown (D_{updown}),
- microsoft/DialogRPT-depth (D_{depth}),
- microsoft/DialogRPT-width (D_{width}),
- microsoft/DialogRPT-human-vs-rand ($D_{\text{vs-random}}$),
- microsoft/DialogRPT-human-vs-machine ($D_{\text{vs-machine}}$).

Table 6 contains the number of responses that came from each of the prompts in the submitted answers to the test set.

4 Results

Despite never using the reference sentences for training or fine-tuning, our system received the highest BERTScore and second highest DialogRPT score on the evaluation data, giving us the highest average rank in the automated metrics.

---

[2] https://github.com/rominf/profanity-filter

[3] All DialogRPT models can be found here: https://github.com/golsun/DialogRPT

	teacher-like	understanding	helpful
NAISTeacher	**2.16**	**2.07**	**1.87**
Reference	3.11	3.10	3.09

Table 7: The average ranking results from the human evaluation. The best possible score is 1 and worst is 4.

The human evaluation of the top three teams was carried out on the Prolific crowdsourcing platform, where our system was compared against the other two systems as well as against the reference teacher utterances. The raters chose the best response on three criteria: (1) which was more likely said by a teacher, (2) which shows a better understanding of the student, and (3) which helps the student more. Table 7 contains the comparison of our system and the reference.

Our system scored the highest out of the three manually evaluated systems on all three criteria, and especially highly on the helpfulness criteria.

5 Discussion

5.1 Alternatives explored

We experimented with several different approaches, including prompt engineering with `GPT-3.5-turbo` (Brown et al., 2020), zero-shot BlenderBot (Roller et al., 2021), and fine-tuning `GPT-2` (Radford et al., 2019). We found that while fine-tuning `GPT-2` led to higher BERTScores, prompting `GPT-3.5-turbo` led to higher DialogRPT scores and sounded much more like a teacher upon manual inspection. Tack and Piech (2022) found that without fine-tuning or prompt tuning, BlenderBot was more successful than `GPT-3` on this task. However, we discovered that with more specific prompting, the output of `GPT-3.5-turbo` could be influenced, so our final system only uses this.

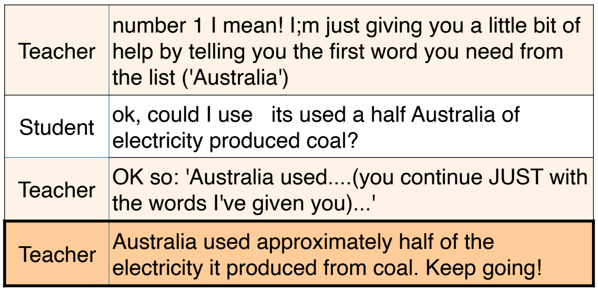
5.2 DialogRPT Length Bias Investigation

One of the drawbacks of using $k = 3$ prompts was that they scored lower on the automated metrics than their $k = 1$ counterparts on average. We inspected the output manually, and on average preferred the responses from the $k = 3$.

To investigate if the lower performance of $k = 3$ prompts was due to a bias in DialogRPT towards longer answers, we added a requested length to two targeted zero-shot prompts: (6) and (8).

(13) *Concatenation of prompt (6) and the following:*

Each response can be 3-5 sentences.



Teacher	number 1 I mean! I;m just giving you a little bit of help by telling you the first word you need from the list ('Australia')
Student	ok, could I use its used a half Australia of electricity produced coal?
Teacher	OK so: 'Australia used....(you continue JUST with the words I've given you)...'
Teacher	Australia used approximately half of the electricity it produced from coal. Keep going!

Figure 3: An example of a teacher continuation that gives the answer instead of scaffolding for the student.

(14) *Concatenation of prompt (8) and the following:*

Each response can be 3-5 sentences.

By adding the length specification of 3-5 sentences, the responses generated by the model become longer, but scored no higher on DialogRPT, leading us to conclude there is no evidence of bias towards longer answers. See Table 8 for detailed results.

5.3 Common errors

An illustration of a common remaining error can be seen in Figure 3. The student has yet to attempt to complete the exercise, and instead of scaffolding the student's attempt to complete it, the teacher responds with the answer.

This could be fixed with a better reranking algorithm, as DialogRPT often scores the responses that contain the correct answer higher than those that scaffold the learner.

For example, the chosen response scored 68.17, while an alternative response that scaffolds better scored 50.34: `Let’s focus on the structure of the sentence next. Remember to use the correct verb form after "Australia used". Also, instead of "half Australia", we would say "half of Australia’s". Could you try revising your sentence to reflect these changes?`

6 Conclusion

The reranker we built was very simple. It selected the highest-scoring response according to the automated metric, DialogRPT. However, preliminary manual evaluation did not always align with DialogRPT. The metric often prefers complete answers that do not encourage student engagement over responses that aim to help the student answer the question for themselves.

	avg. characters	avg. words	BERTScore	DialogRPT
(1) Zero-shot, $k = 1$	205.16	35.55	70.91	36.51
(1) Zero-shot, $k = 3$	137.09	23.67	70.04	36.27
(6) Reply, $k = 3$	170.76	29.67	83.29	36.39
(13) Reply long, $k = 3$	231.58	40.77	70.82	34.68
(8) Continue, $k = 3$	145.43	25.52	70.09	36.29
(14) Continue long, $k = 3$	238.71	42.26	69.49	34.76
Reference	126.26	23.01	1.00	32.81

Table 8: The average length of the responses of prompts with $k = 1$ and $k = 3$, as well as those with a length of 3-5 sentences specified in the prompt.

With additional time, we would like to develop a model capable of classifying the extent to which a generated response reflects a teacher's style. This model could take into account whether the response effectively balances helpfulness with scaffolding independent thought, as well as the degree to which it demonstrates an understanding of the student's needs. Such a model would lead to improved performance of the reranker.

The AI teacher response generator we created still needs improvement before it can become a fully functional teacher chatbot. The responses it generates can be excessively detailed at times due to the automated metrics used, which prioritize comprehensive responses. When integrated into an assistant, it may seem as if the responses are repetitive or that there is no well-designed lesson plan in place.

As we were building the system, we kept in mind how it was to be evaluated. That is, by machine first checking for similarity to a reference answer as well as usefulness and relevance to the conversation, then by humans evaluating how teacher like, understanding, and helpful the response was. We used a combination of $k = 1$ prompts which scored higher on the automated metrics, and $k = 3$ prompts which produced shorter responses that we preferred on manual evaluation.

While the automated metrics taken together align with human evaluators' judgments, DialogRPT alone does not always correspond with human judgment. The DialogRPT score models were trained on Reddit, which follows a different format than live chat. Reddit is an asynchronous format, meaning that it tends to have longer, more complete responses. On the other hand, synchronous chat-based lessons feature multiple consecutive turns, as it is more common in instant messaging to break up longer thoughts into smaller turns. DialogRPT was not trained to judge the continuation of a response, which made it less reliable as a reranker for teacher continuations, in particular.

When chatting, it's not necessary to always respond with a detailed message. The reference teacher responses offer a mix of quick replies, corrections, elaborations, practice activities, and clarifications, among others. In the future, we would like to incorporate more of the conversational moves that real teachers use in these types of exchanges.

In conclusion, our approach to generating teacher utterances in an educational dialogue for the BEA2023 shared task used a pretrained language model and an ensemble of prompts to generate multiple candidates, which we then reranked using automated metrics. We experimented with different techniques for generating few-shot prompts and found that an iterative approach was the most effective. Our system achieved the highest averaged ranked scores in both the automated and human evaluation rounds. Overall, our approach shows promise for generating effective and helpful teacher utterances in educational dialogues.

Limitations

A limitation of our approach is that it relies heavily on the quality and relevance of the prompts used. The prompts were engineered based on observations made in the training data and this approach may not work if the prompts are not representative of the corpus. Finally, our approach may not be suitable for all types of teacher-student dialogues and may require modifications for different contexts or domains.

One possible concern with the techniques mentioned in this paper is the limited reproducibility of OpenAI's language models, such as `GPT-3.5-turbo`. The weights of these models are proprietary and not publicly accessible, which makes it challenging to replicate the findings of earlier research or expand on them.

Ethical Considerations

AI-generated teacher utterances may contain bias, which may become apparent particularly in exercises or chit-chat.

In this project, we took steps to avoid profanity in the AI-generated responses, but similar protection against bias should be put into place. Additionally, human evaluators should be used to assess the quality of the AI-generated responses and to identify any potential biases.

We recognize that language models like `GPT-3.5-turbo` are trained on large datasets that reflect the biases and prejudices present in society. As there is always a risk of perpetuating these biases when using generative AI for dialogue systems it is important to evaluate the AI-generated responses for potential biases and to take steps to correct them.

Acknowledgements

We are very grateful for the thoughtful and helpful feedback from our reviewers, and to Xincan Feng, Iqra Ali, and Hidetaka Kamigato for their support and feedback.

References

Dennis Aumiller and Michael Gertz. 2022. UniHD at TSAR-2022 shared task: Is compute all we need for lexical simplification? In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 251–258, Abu Dhabi, United Arab Emirates (Virtual). 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. *CoRR*, abs/2005.14165.

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.

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

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.

Ann Lee, Michael Auli, and Marc'Aurelio Ranzato. 2021. Discriminative reranking for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7250–7264, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *CoRR*, abs/1510.03055.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 295–302, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Libin Shen, Anoop Sarkar, and Franz Josef Och. 2004. Discriminative reranking for machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 177–184, Boston, Massachusetts, USA. Association for Computational Linguistics.

Anaïs Tack, Ekaterina Kochmar, Zheng Yuan, Serge Bibauw, and Chris Piech. 2023. The BEA 2023 Shared Task on Generating AI Teacher Responses in Educational Dialogues. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building*

Educational Applications, page to appear, Toronto, Canada. Association for Computational Linguistics.

Anaïs Tack and Chris Piech. 2022. The AI Teacher Test: Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues. In *The 15th International Conference on Educational Data Mining*, page accepted.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

A Zero-shot iterations

The zero-shot prompt went through several iterations. We present six versions along with the responses for the following example conversation:

> Teacher: keywords they focus on in that page i mean
> Student: OK, I'll do it now
> Teacher: ok
> Student: ...provocative, and supplied +startling+ messages about psycology...
> Student: Is it enough this piece of text to work on it? the word I don't konw is startling
> Teacher: ok thanks....'startling' does it mean: obvoius or surprising do you think? is there a context in the text to help you?

(15) Can you give an example teacher response that would be helpful for the language learner and encourage the continuation of the conversation?

> NAISTeacher: Great job providing those keywords! Let's take a closer look at the word "startling." It means something that is very surprising or shocking. For example, "The news of the unexpected victory was startling to everyone." In the context of the text you provided, it seems that the messages about psychology were surprising and caught your attention. Keep up the good work! Is there anything else you're unsure about or need help with?

The response from prompt (15) on this example reveals several potential problems:

- It answered the teacher's question before allowing the student to reply.
- It helpfully provides an example sentence, but the inclusion of this makes the reply unnaturally long.
- It sounds as though the teacher is unfamiliar with the text they assigned.
- The final question sounds like an assistant or a chatbot, not a teacher. While it is helpful and pleasant, it lacks a sense of agency or authority in the situation.

With these problems in mind, we iterated on the prompt.

(16) Can you give an example teacher **follow-up to their previous message** that would be helpful for the language learner ~~and encourage the continuation of the conversation~~?

> NAISTeacher: Good job identifying those keywords! 'Startling' means surprising or shocking. In this context, it suggests that the messages about psychology were not only provocative, but also had an element of surprise or shock to them. Keep up the good work!

With the exclusion of `and encourage the continuation of the conversation` the response ends too finally. We want a way to keep the student engaged and active in the conversation.

(17) Can you give an example teacher follow-up to their previous message that would be helpful for the language learner? **The message should encourage the continuation of the current topic, or gracefully transition to a new topic.**

> NAISTeacher: Great job using the keywords from the text, it seems like you have a good understanding of the content. To answer your question, 'startling' means surprising or unexpected. In the context of the text, it suggests that the messages about psychology were particularly surprising or unexpected. Keep up the good work! Is there anything else you'd like to discuss related to this text or another topic?

The end of the response improved, but is too open ended, and sounds unnatural for a teacher.

(18) Can you give an example teacher follow-up to their previous message that would be helpful for the language learner? The message should encourage the continuation of the current topic, or gracefully transition to a new teacher-provided topic. **Try to not sound like an assistant, but a teacher, in charge of the flow of the lesson.**

> NAISTeacher: Great job identifying those keywords! To answer your question, 'startling' means surprising or shocking. It's often used to describe something that catches you off guard or is unexpected. In this context, it seems like the article is trying to grab the reader's attention with provocative and surprising messages about psychology.
>
> Now that we've clarified the meaning of 'startling,' let's continue with our analysis of the text. Can you tell me more about what you think the article is trying to convey? Do you have any questions or ideas about the content?

The question at the end is much more specific and teacher-like, but there remains the problem of the teacher over-explaining and answering a question too directly. We wanted the teacher to encourage the student to think rather than just read an explanation.

(19) Can you give an example teacher follow-up to their previous message that would be helpful for the language learner? **Instead of answering the student's question directly, give hints and encourage the student to think for themselves.** The message should encourage the continuation of the current topic, or gracefully transition to a new teacher-provided topic. Try to not sound like an assistant, but a teacher, in charge of the flow of the lesson.

> NAISTeacher: Great job, you're making progress! You're right that "startling" is not a very common word, but it's definitely a useful one to know. To answer your question, "startling" means surprising, or even shocking. So when you see that word in a text, it usually means that something unexpected or surprising is being described. In this context, it seems that the page you were reading contained messages about psychology that were not only

> provocative, but also surprising or shocking in some way.
>
> It's always a good idea to try to understand the context of a word in order to better understand its meaning. Sometimes, the words that come before or after it can help you understand what it means. For example, in this case, the word "provocative" suggests that the messages were intended to be controversial or challenging, and "startling" adds to this by suggesting that the messages were even more unexpected than one might expect from a provocative message.
>
> Keep up the good work, and don't hesitate to ask if you have any more questions or if you want to practice using these words in conversation or writing!

The response to this prompt is way too long, and the inclusion of `Instead of answering the student's question directly, give hints and encourage the student to think for themselves` did not solve the problem. We remove that and add the stipulation that the response be concise.

(20) Can you give an example teacher follow-up to their previous message that would be helpful for the language learner? **The message should be concise, and worded simply.** ~~Instead of answering the student's question directly, give hints and encourage the student to think for themselves.~~ It should either encourage the continuation of the current topic or gracefully transition to a new teacher-provided topic. **Questions should be specific and not open-ended.** Try to not sound like an assistant, but a teacher, in charge of the flow of the lesson.

> Great job using the keywords from the text! "Startling" means surprising. In this context, it seems like the text contains provocative and surprising messages about psychology. Can you think of any other words that might be related to the topic of psychology?

We found this to be a reasonable response. It still has the problem of answering the question posed by the teacher, which we chose to address by splitting out teacher continuations (see 3.4.3 Specific teacher continuation scenarios), but the response is a manageable length and invites the student to engage with the conversation further and think about the topic in more depth.