The BEA 2023 Shared Task on Generating AI Teacher Responses in Educational Dialogues

Anaïs Tack
KU Leuven, imec
anais.tack@kuleuven.be

Ekaterina Kochmar
MBZUAI
ekaterina.kochmar@mbzuai.ac.ae

Zheng Yuan
King's College London
zheng.yuan@kcl.ac.uk

Serge Bibauw
Universidad Central del Ecuador
sbibauw@uce.edu.ec

Chris Piech
Stanford University
cpiech@stanford.edu

Abstract

This paper describes the results of the first shared task on generation of teacher responses in educational dialogues. The goal of the task was to benchmark the ability of generative language models to act as AI teachers, replying to a student in a teacher–student dialogue. Eight teams participated in the competition hosted on CodaLab and experimented with a wide variety of state-of-the-art models, including Alpaca, Bloom, DialoGPT, DistilGPT-2, Flan-T5, GPT-2, GPT-3, GPT-4, LLaMA, OPT-2.7B, and T5-base. Their submissions were automatically scored using BERTScore and DialogRPT metrics, and the top three among them were further manually evaluated in terms of pedagogical ability based on Tack and Piech (2022). The NAISTeacher system, which ranked first in both automated and human evaluation, generated responses with GPT-3.5 Turbo using an ensemble of prompts and DialogRPT-based ranking of responses for given dialogue contexts. Despite promising achievements of the participating teams, the results also highlight the need for evaluation metrics better suited to educational contexts.

1 Introduction

Conversational AI offers promising opportunities for education. Chatbots can fulfill various roles – from intelligent tutors to service-oriented assistants – and pursue different objectives such as improving student skills and increasing instructional efficiency (Wollny et al., 2021). One of the most important roles for an educational chatbot is that of an AI teacher which helps a student improve their skills and provides more opportunities to practice. Recent studies suggest that chatbots have a significant effect on skill improvement, for example, in language learning (Bibauw et al., 2022). Moreover, the advances in Large Language Models (LLMs) open up new opportunities as such models have a potential to revolutionize education and significantly transform learning and teaching experience.

Despite these promising opportunities, the use of powerful generative models as a foundation for downstream tasks presents several crucial challenges, in particular, when such tasks may have real social impact. Specifically, in the educational domain, it is important to determine how solid that foundation is. Bommasani et al. (2021) (pp. 67-72) stresses that if we want to put such models into practice as AI teachers, it is of crucial importance to determine whether they can (a) speak to students like a teacher, (b) understand students, and (c) help students improve their understanding. Following these desiderata, Tack and Piech (2022) formulated the AI teacher test challenge: *How can we test whether state-of-the-art generative models are good AI teachers, capable of replying to a student in an educational dialogue?*

Building on the AI teacher test challenge, we have organized the first shared task on generation of teacher language in educational dialogues. The goal of this task is to explore the potential of NLP and AI methods in generating teacher responses in the context of real-world teacher–student interactions. Interaction samples werer extracted from the *Teacher Student Chatroom Corpus* (Caines et al., 2020, 2022), with each training sample consisting of a dialogue context (i.e., several rounds of teacher-student utterances) and the teacher's response. For each test sample, participants were asked to submit their best generated teacher response.

As the purpose of this task was to benchmark the ability of generative models to act as AI teachers, responding to a student in a teacher–student dialogue, submissions were first ranked according to popular BERTScore and DialogRPT metrics, and the top three submissions were then selected for further human evaluation. During this manual evaluation, the raters compared a pair of "teacher" responses along three dimensions: speaking like a teacher, understanding a student, and helping a student (Tack and Piech, 2022).

Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 785–795

SPEAKER	UTTERANCE	
Teacher:	Yes, good! And to charge it up, you need to __ it ___	DIALOGUE CONTEXT
Student:	…	
Teacher:	connect to the source of electricity	
Student:	i understand	
Teacher:	plug it __?	
Student:	in	
Teacher:	yes, good. And when the battery is full, you need to ____ (disconnect it)	= REFERENCE RESPONSE

Figure 1: An example of a sample taken from the *Teacher-Student Chatroom Corpus*

2 Materials and Methods

The shared task used data from the *Teacher-Student Chatroom Corpus* (TSCC) (Caines et al., 2020, 2022). This corpus comprises data from several chatrooms in which an English as a second language (ESL) teacher interacts with a student in order to work on a language learning exercise and assess the student's English language proficiency.

2.1 Data Samples

Several samples were taken from each dialogue in the corpus. Each sample was composed of several sequential teacher-student turns (i.e., the preceding dialogue context) and ended with a teacher utterance (i.e., the reference response). Figure 1 shows an example of a sample taken from the corpus. As can be seen from this example, the samples were quite short, counting at most 100 tokens. Even though this restricted sample size inevitably posed an important limitation for training and testing, the length of each sample had to be capped at this specific limit in order to comply with the copyright license and terms of use of the corpus.

2.1.1 Extraction

The samples were extracted with the following method. For each dialogue in the corpus, the sequence of utterances was iterated from the first to the last. If the speaker of an utterance at the current position was a teacher, the utterance was a potential reference response. In that case, a contextual window sequence was created for the reference candidate by recursively backtracking through the dialogue and adding the preceding utterances until the limit of 100 tokens was reached. Each utterance was tokenized with spaCy's default tokenizer for English.[1] Once extracted, the sequence was added to the set of samples for the dialogue on the condition that it had at least two utterances and more than one speaker. For example, if the teacher initiated the conversation, the algorithm would extract a window with only one speaker and no preceding utterances. Because this instance would not have been informative, it was ignored and not added to the set of data samples. A total of $7,047$ data samples were extracted from the original dataset.

2.1.2 Selection

Although the extracted data samples could have been randomly divided into training and test samples, such an approach would have been problematic. In fact, it would have been possible for a randomly selected test sample to contain a reference response otherwise observed in the dialog context of *another* randomly selected training or test sample (see Figure 2). A related issue was that the extraction algorithm produced samples that were also part of other samples, resulting in multiple nested or Russian doll-like ensembles (see Figure 3). Since a test set should never include references seen elsewhere in the data, special attention was paid to data splitting.

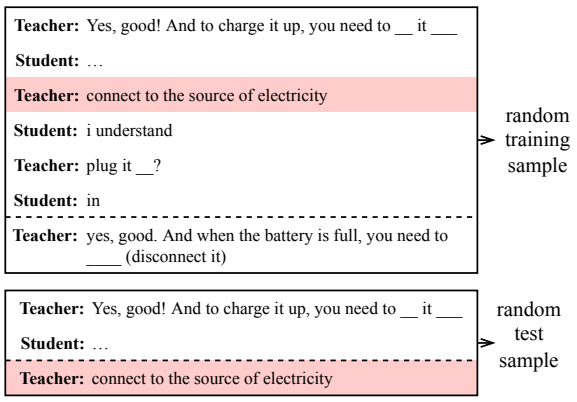

Figure 2: An example of a reference in a test sample observed in the context of a training sample

[1] https://spacy.io/api/tokenizer

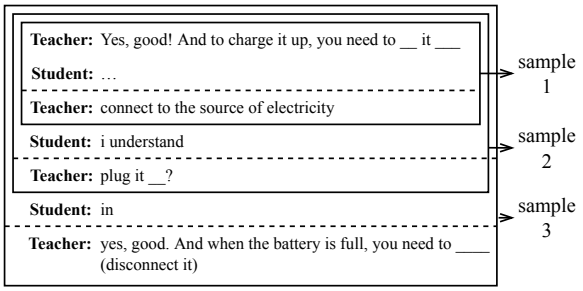

Figure 3: An example of a nested or Russian doll-like ensemble of data samples

The data samples were split into a training and test set with a more complex selection procedure. Three selection criteria were defined: (a) whether the reference response was labeled as *eliciting* and/or *scaffolding* ('yes' ⇒ better), (b) the number of distinct types of conversational organization (e.g., opening, closing, eliciting, scaffolding, and revision) that were added as labels to the reference response (more ⇒ better), and (c) the total number of tokens in the sample (more ⇒ better). The extracted data samples contained $1,400$ nested ensembles (cf. Figure 3). The samples in each ensemble were sorted based on the three criteria above, and for each ensemble, only the best sample was selected. The remaining $4,864$ samples were assigned to $2,457$ training and 273 test slots with the Hungarian algorithm (Kuhn, 1955) based on the criteria above. Once the assignment was done, the training and test sets were verified for any potential conflicts (cf. Figure 2). Conflicts were resolved by using the criteria above to choose the best sample among the conflicting samples. Then, the assignment was run again on the remaining samples until no more conflicts could be detected. After the assignment was completed, the nested data samples that were discarded before were used to increase the size of the training set on the condition that they were not in conflict with the test set. Finally, the training set was randomly split into a 90% training and 10% held-out set. The number of samples included in the training and test sets are shown in Table 1.

Training set	$3,052$	
	90% training	$2,747$
	10% held-out	305
Test set	273	

Table 1: The number of training and test samples

2.2 Competition

The shared task was hosted as an online competition on the CodaLab platform (Pavao et al., 2022). Anyone participating in the shared task filled in a registration form, signed to comply with the terms and conditions of the shared task and the licensed TSCC data, and registered on the CodaLab platform. Participants could only be part of one team, while a team could have one or more participants.

2.2.1 Phases

The competition was run in two phases: a development and an evaluation phase. All deadlines were set to 23:59 Anywhere on Earth (UTC-12). Since CodaLab uses Coordinated Universal Time, all deadlines on the platform were adapted accordingly (i.e., set to the next day at 11:59 am UTC).

The development phase started on March 24, 2023, and ended on April 30, 2023. At the start of the development phase, participants received the training and held-out development data, which were available on the CodaLab platform. During the development phase, participants could submit their results for the held-out data and view their scores on the anonymized leaderboard. Sixty-three people filled in the registration form and registered on the CodaLab platform. Among them, 12 people actively participated in the development phase and submitted results on the held-out data. Three people submitted to the development phase after the evaluation phase had already started. In the end, 10 participants made at least one successful submission to the development phase. In total, 17 successful submissions were received ($M_{\text{submissions}} = 1.7$ per participant). The leaderboard featured only the best successful submission per participant (see the metrics described below in Section 2.3.1).

The evaluation phase started on May 1st, 2023, and ended on May 5th, 2023. At the start of the evaluation phase, participants received the test data, which were available on the CodaLab platform. During the evaluation phase, participants could submit their results on the test data and view their scores on the anonymized leaderboard. In addition, six people filled in the registration form and registered on the CodaLab platform. Nineteen people actively participated in the evaluation phase and submitted their results on the test data. In the end, 10 participants from eight teams made at least one successful submission to the evaluation phase. In total, 19 successful submissions were received

($M_{\text{submissions}} = 1.9$ per participant). Again, the leaderboard featured only the best successful submission per participant (see the metrics described in Section 2.3.1).

It should be noted that some people showed interest in the shared task but did not fully participate. Fifteen people filled in the registration form but did not request to join on the platform before the deadline, whereas 18 people requested to join on CodaLab but did not fill in the registration form. As a result, they could not be accepted into the competition because they did not sign to comply with the terms and conditions.

2.2.2 Teams and Systems

Eight teams made at least one successful submission to the final evaluation phase. The approaches taken by the teams were based on a range of state-of-the-art large language models (LLMs), including Alpaca (Team RETUYT-InCo), Bloom (RETUYT-InCo), DialoGPT (Cornell), DistilGPT-2 (DT), Flan-T5 (teams Cornell and TanTanLabs), GPT-2 (Cornell and Data Science-NLP-HSG), GPT-3 (NBU), GPT-3.5 Turbo (NAIST and aiitis), GPT-4 (Cornell), LLaMA (RETUYT-InCo), OPT-2.7B (RETUYT-InCo), and T5-base (Data Science-NLP-HSG). In addition, all teams experimented with zero- and few-shot learning, fine-tuning, and various prompting strategies. Several teams applied reinforcement learning (RL) (Cornell and Data Science-NLP-HSG), and some developed customized approaches to post-processing (NAIST) and data-driven prompt engineering (aiitis). All these approaches are summarized below and further detailed in the corresponding system papers.

Team NAIST Vasselli et al. (2023) participated in the shared task with the NAISTEACHER system, built on a pre-trained GPT-3.5 Turbo (Brown et al., 2020). They experimented with, on the one hand, zero-shot prompts and, on the other hand, few-shot prompts using either handcrafted, generative, or iterative examples of teacher responses. They also experimented with asking the model to generate either one response or several possible responses and compared the performance of their system in two settings: *teacher replies* (i.e., when the generated teacher utterance followed a student utterance) and *teacher continuations* (i.e., when the generated teacher utterance followed a teacher utterance). Finally, the candidate responses were post-processed (with a profanity filter and regular expressions) and reranked with DialogRPT (see the shared task metrics in Section 2.3.1) in order to select the best response to be submitted for each test sample.

Team NBU Adigwe and Yuan (2023) participated in the shared task with the ADAIO system. They evaluated several GPT-3 models (Brown et al., 2020), designed various zero-shot and few-shot prompts to generate teacher responses, and also fine-tuned the models on the TSCC corpus. In addition, the team experimented extensively with various aspects of response generation by considering the roles of the participants, the teaching approaches taken by the tutor, and the specific teaching goals. The responses submitted to the competition were generated by a few-shot prompt-based method based on the *text-davinci-003* model.

Team Cornell Hicke et al. (2023) experimented with several generative models and various approaches, including few-shot in-context learning with GPT-4, fine-tuning of GPT-2 (Radford et al., 2019) and DialoGPT (Zhang et al., 2019), and fine-tuning of Flan-T5 (Chung et al., 2022) with RL (Ramamurthy et al., 2022) to optimize for pedagogical quality. Among these, GPT-4 achieved the best results on the shared task evaluation metrics (see Section 2.3.1). The team made two submissions to the leaderboard: one submission with responses generated by GPT-4, and another submission that included the same responses with a teacher prefix prepended to each of them (`"teacher: <response>"`). To distinguish between these submissions, the latter is referred to as GPT-4 $^{(\text{TP})}$ where TP stands for teacher prefix.

Team aiitis Omidvar and An (2023) introduced the Semantic In-Context Learning (S-ICL) model. Their aim was to address the challenges created by the use of out-of-the-box pre-trained LLMs, such as domain adaptivity and the high costs of fine-tuning. Their in-context learning approach consisted of providing an LLM (in this case, ChatGPT with the GPT-3.5 Turbo engine) with a prompt containing an instruction, a few labeled samples, and an unlabeled sample. The *semantic* component in the S-ICL model retrieved sufficiently similar samples from the training set, which were then integrated into the prompt fed to the LLM as labeled samples. The inclusion of relevant conversational samples in the prompt allowed the model to leverage available knowledge for generating teacher responses.

Team RETUYT-InCo Baladón et al. (2023) experimented with several open-source LLMs, including LLaMA (Touvron et al., 2023), Alpaca (Taori et al., 2023), OPT-2.7B (Gao et al., 2020a), and Bloom 3b (Scao et al., 2022). They explored fine-tuning techniques by applying the LoRA (Hu et al., 2021) method to the aforementioned LLMs. They tested several prompting strategies including few-shot and chain-of-thought approaches. Their method consisted of selecting the three most similar conversations from the training data using the k-nearest neighbors algorithm. These were then further integrated into the prompt for the few-shot learning scenario. The models submitted to the competition were trained using Alpaca LoRA with the few-shot approach, LLaMA 7B with engineered prompts fine-tuned with LoRA, and fine-tuned OPT-2.7B using preprocessing.

Team Data Science-NLP-HSG Huber et al. (2023) presented a simple approach of fine-tuning a language model with RL and utilized the novel NLPO algorithm (Ramamurthy et al., 2022) that masks out tokens during inference to direct the model towards generations that maximize a reward function. They used Hugging Face's implementation of the T5-base model (Raffel et al., 2020) with 220 million parameters to generate the responses submitted to the competition.

Team DT This team experimented with fine-tuning the DistilGPT-2 model specifically for student–teacher dialogues. They divided the original training data using an 80/20 split and ran a three-epoch training process using the Adam optimizer along with a linear learning rate scheduler on the training subset. The remaining 20% were then used for rigorous evaluation using the shared task performance metrics. The team released their model on Hugging Face and plans to explore the potential of larger models like GPT-3 and GPT-4 in the educational dialogue domain in the future.[2]

Team TanTanLabs This team experimented with a zero-shot approach using Hugging Face's Flan-T5 transformer model, a model instruction-finetuned on a mixture of tasks. Among the many prompting techniques tested, the one that worked best was the prompt used by the authors of the Flan-T5 model: "Read the dialog and predict the next turn." For model inference, different decoding techniques were tried (greedy, decoding by sampling with temperature, and beam search). Beam search was chosen because it was easy to control. Customized regular expressions were used to parse the model's output. When the model didn't produce any output, the filler word "Alright" was used. In the future, the team plans to experiment further with supervised fine-tuning using "chain of thought" reasoning instructions.[3]

2.3 Evaluation Procedure

The submissions made by the teams described above were evaluated in two stages. During the competition, all submissions were automatically scored with several dialogue evaluation metrics (see Yeh et al., 2021, for a comprehensive review). The teams used these metrics to optimize their systems before the end of the competition. After the competition ended, the final submissions were evaluated by human raters. Due to combinatorial constraints imposed by the human evaluation task (see Section 2.3.2), it was not possible for any number of submissions to be evaluated manually. For this reason, only the top three submissions on the automated metrics were targeted for human evaluation.

2.3.1 Evaluation Metrics

Yeh et al. (2021) reviewed several dialogue evaluation metrics that operate at the level of the individual turns (i.e., generated responses). However, many of these metrics required a complicated installation procedure. The following two metrics were used because they are well-known, could be easily installed, and their scores can be reproduced.

BERTScore (Zhang et al., 2020) was used as a metric for evaluating each generated response with respect to the reference (i.e., teacher) response. The metric matches words in submissions and reference responses by cosine similarity. BERTScore was computed with Hugging Face's *evaluate* package and the *distilbert-base-uncased*[4] model. The resulting precision, recall, and F1 scores were averaged for all items in the test set.

DialogRPT (Gao et al., 2020b) was used as a reference-free metric for evaluating the generated response with respect to the preceding dialogue context. The metric consists of a set of ranked pre-trained transformer models proposed by Microsoft

[2]Written by Rabin Banjade and adapted by the authors

[3]Written by Tanay Gahlot and adapted by the authors

[4]The hashcode was *distilbert-base-uncased_L5_no-idf_version=0.3.12(hug_trans=4.28.1)*.

Research NLP Group. These metrics were aggregated for all items in the test set. The following dialog response ranking models were used:

updown likelihood that a response gets the most upvotes (mean of all items)

human vs. rand likelihood that a response is relevant for the given context (mean of all items)

human vs. machine likelihood that a response is human-written rather than machine-generated (mean of all test items)

final weighted ensemble score of all DialogRPT metrics (mean of all items)

Each submission was ranked from 1 (highest) to 10 (lowest) on each individual metric. The overall leaderboard rank was computed as the mean rank on BERTScore F1 and on DialogRPT final average. In case of a tie, the tiebreaker was the mean rank on the individual scores for BERTScore (precision, recall) and DialogRPT (updown, human vs. rand, human vs. machine).

2.3.2 Human Evaluation

The top $k = 3$ submissions on the leaderboard were further evaluated by means of pairwise comparative judgments.[5] For each sample in the set of $n = 273$ test items, the possible responses were combined in pairs such that the generated responses were either compared with the reference (i.e., teacher vs. AI) or between themselves (i.e., AI vs. AI). This resulted in $\binom{k+1}{2} = 6$ pairs of responses for each test sample. Each pair was assessed by $r = 3$ raters, which amounted to a total of $\frac{(k+1)!}{2!(k+1-2)!}r = 4,914$ distinct assessments. These evaluations were collected via an online Qualtrics survey following a method described in Tack and Piech (2022) and further detailed below.

Survey In the introductory part of the survey, raters were given a short introduction, a consent form, and an example to familiarize themselves with the task at hand. In the central part of the survey, each rater was presented with a comparative judgment task of 20 items that were randomly and evenly selected from the set of n test samples. Each survey item included a pairwise comparison that was randomly and evenly selected from the $\binom{k+1}{2}$ possible pairs for the chosen test sample. Each survey item had three components: the dialogue context, one comparison of two responses (A or B), and three questions targeting a pedagogic ability (*more likely said by a teacher*, *better understanding the student*, and *helping the student more*). For each question, the rater was asked to choose option A or B. The order in which the pairwise comparison was presented, was determined randomly so that any presentation order effects would be avoided.

Raters A sample of 298 raters were recruited from the Prolific crowdsourcing platform. The raters were screened based on several characteristics: (a) whether they were from a majority native English-speaking country,[6] (b) whether their native language was English, and (c) whether their employment sector was in education and training. The sample of raters was gender-balanced. Five raters were removed because the outlier detection described in Tack and Piech (2022) showed that they consistently picked the same option (A or B) for all questions throughout the survey.

Ranking For each item in the test set, the possible responses were ranked from 1 (highest) to 4 (lowest) for each of the three questions (*more likely said by a teacher*, *understanding the student better*, and *helping the student more*). The rank for each response (i.e., teacher or AI) was estimated with a Bayesian Bradley-Terry model and HMC-NUTS sampler as described in Tack and Piech (2022). Based on the set of draws produced by the HMC-NUTS sampler, the mean rank, standard deviation, and 95% highest density intervals (HDI) were computed for each item and for each response.

3 Results

The results achieved by the participating teams during the automated evaluation phase are shown in Table 2 and those achieved by the top three during the human evaluation phase are shown in Figure 4.

As can be observed from Table 2, the NAIS-Teacher system (Vasselli et al., 2023) attained the highest average rank on BERTScore and DialogRPT. On average, the responses were the closest to the teacher's response, the most relevant for the

[5] In pairwise comparative judgments, multiple alternatives are evaluated by systematically assessing them in pairs. Each rater is presented with two alternatives at a time and makes a judgment about which one is better according to some criteria. These judgments are used to compute a relative ranking among the alternatives. This method has already been used for assessing dialogue systems (Li et al., 2019) and open-ended natural language generation (Pillutla et al., 2021).

[6] Based on the UK government classification + Ireland.

Team	System	BERTScore			DialogRPT				Rank
		P	R	**F1**	U	HvR	HvM	**Final**	
NAIST	NAISTeacher	0.71 (9)	**0.71** (1)	**0.71** (1)	0.48 (2)	**0.98** (1)	**1.00** (1)	0.46 (2)	1.5
NBU	ADAIO	0.72 (4)	0.69 (3)	0.71 (3)	0.40 (5)	0.97 (2)	0.98 (5)	0.37 (3)	3.0
Cornell	GPT-4 (TP)	0.71 (7)	0.69 (2)	0.70 (5)	**0.52** (1)	0.86 (8)	0.98 (2)	**0.47** (1)	3.0
aiitis	S-ICL	0.72 (3)	0.69 (5)	0.70 (4)	0.40 (4)	0.92 (5)	0.98 (4)	0.36 (5)	4.5
RETUYT-InCo	OPT-2.7B	**0.74** (1)	0.68 (6)	0.71 (2)	0.38 (7)	0.90 (7)	0.96 (9)	0.35 (7)	4.5
Cornell	GPT-4	0.72 (5)	0.69 (4)	0.70 (6)	0.40 (6)	0.93 (4)	0.98 (3)	0.36 (6)	6.0
Data Science-NLP-HSG	Untrained	0.72 (6)	0.63 (8)	0.67 (8)	0.41 (3)	0.93 (3)	0.95 (10)	0.37 (4)	6.0
RETUYT-InCo	Alpaca	0.72 (2)	0.68 (7)	0.70 (7)	0.37 (8)	0.91 (6)	0.96 (7)	0.34 (8)	7.5
DT	DistilGPT2	0.67 (10)	0.62 (9)	0.64 (10)	0.36 (9)	0.75 (10)	0.96 (6)	0.29 (9)	9.5
TanTanLabs	zero-shot-with-filler	0.71 (8)	0.60 (10)	0.65 (9)	0.32 (10)	0.85 (9)	0.96 (8)	0.29 (10)	9.5
TEACHER	REFERENCE	1.00	1.00	1.00	0.37	0.86	0.99	0.32	

Table 2: Leaderboard for the evaluation phase with scores and ranks for BERTScore (P = precision, R = recall) and DialogRPT (U = updown, HvR = human vs. rand, HvM = human vs. machine)

given dialogue context, and also the most likely to be human-written. The system also achieved the second-best result on the DialogRPT updown metric, which indicated that the generated responses were likely to receive upvotes. Besides achieving the best average rank on the evaluation metrics, the system also achieved the best rank on all three criteria of pedagogical ability evaluated by human raters (see Figure 4). In particular, the responses were found to be the most helpful overall.

Table 2 further shows that the best result on the DialogRPT updown metric was achieved by the Cornell team (Hicke et al., 2023). The responses generated by GPT-4 were the most likely to receive upvotes on average (0.52) when they were submitted with a teacher prefix. However, when the team submitted the same responses *without* the prefix, they received a much lower score (0.4) and ranked 6th place on the same metric. This remarkable outcome highlighted the unanticipated sensitivity of the DialogRPT metric towards the presence or absence of a prefix.

The ADAIO system (Adigwe and Yuan, 2023) attained the second-best average rank on both the automated evaluation phase (Table 2) and the human evaluation phase (Figure 4). The results indicated that the use of well-engineered prompts including good teaching examples (NAISTeacher, #1) and teaching approaches and goals (ADAIO, #2) resulted in a high rank on BERTScore, DialogRPT, and assessments of pedagogic ability.

It is interesting to note that the teacher's response was ranked *lower* than the top three systems built on GPT-3 and GPT-4 (Figure 4), which contradicts the results of Tack and Piech (2022). This striking observation might be explained by some differences in the human evaluation procedure: while any native English speaker could participate in Tack and Piech (2022), only raters working in education and training could participate in the shared task. Some of these raters gave specific feedback stating that they found the non-standard language used by the teacher in the chatroom (including spelling mistakes, typos, and such) unprofessional.

For more in-depth analyses, the reader is referred to the system papers cited in this paper.

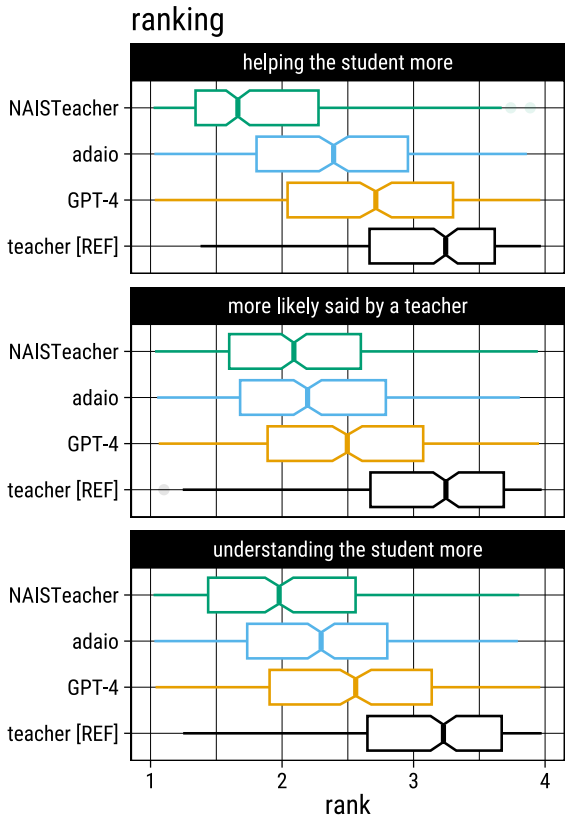

Figure 4: Ranking of the top three submissions and the teacher reference after the human evaluation phase

In these papers, the participating teams ran additional analyses and made critical observations. For example, Baladón et al. (RETUYT-InCo) observed that fine-tuned models attained better results on BERTScore, prompting attained better results on DialogRPT, and methods that combined both techniques showed competitive results across all metrics. At the same time, they found that a baseline generating "Hello" in response to every prompt achieved the best result for BERTScore precision and DialogRPT updown. Huber et al. (Data Science-NLP-HSG) found that GPT-2 – a smaller model with 124 million parameters – achieved competitive performance compared to the T5-base model. Moreover, they found that, even though they maximized BERTScore F1 as a reward function, their model scored highly in terms of the other evaluation metrics. Vasselli et al. (NAIST) noted that DialogRPT often preferred complete answers that were not very teacher-like over responses that helped the student find the answer by themselves.

4 Discussion

Although the inaugural shared task on generating AI teacher responses in educational dialogues can be considered a success, the results demonstrate that the evaluation of natural language generation models remains challenging. Ultimately, we would like to have at our disposal precise, valid, and – ideally – automated methods that reward machines and/or humans for their pedagogical abilities. However, we are probably still a long way from achieving this ultimate goal.

The automated metrics that currently exist are not capable of rewarding models for their ability to showcase pedagogical skills. In particular, to the best of our knowledge, there does not exist any comprehensive metric capable of evaluating whether responses are likely to be produced by a teacher, as well as whether they demonstrate understanding of what the student is saying and are helping the student. Moreover, popular automated metrics such as BERTScore and DialogRPT used in this task show a considerable sensitivity to construct-irrelevant variations, as is demonstrated by the use of a "Hello" baseline (Baladón et al., 2023) and an inclusion of the "teacher:" prefix (Hicke et al., 2023). Future editions of this task should, therefore, aim to either develop or resort to more accurate and domain-specific automated metrics as per the observations and suggestions from several competing teams (Adigwe and Yuan, 2023; Baladón et al., 2023; Hicke et al., 2023; Vasselli et al., 2023).

Due to the lack of adequate metrics, we need to resort to manual evaluation methods in order to achieve more precise assessments. However, a typical drawback to manual evaluation is that it is very costly and time-consuming to have a sufficient number of raters evaluating *any* possible response that can be generated in the large space of possible teacher replies. Due to practical and budgetary limitations, it is challenging to organize a shared task during which any possible number of submissions can in principle be evaluated with adequately remunerated human evaluations.

What is more, data is very important in the context of real-world applications and shared tasks. Although the corpus used in this shared task is a valuable resource in our domain, some particularities of this corpus and the data sampling method also had an undeniable impact on the results. Therefore, in future editions of this shared task we should

rethink some of the current potential limitations, such as the fact that the dialogues had to be limited to 100 tokens, resulting in partial conversations; the fact that some dialogues, if extracted from the data randomly might have led to data leakage; and the fact that the dialogues did not always follow strictly role-alternating format, with some teacher turns being preceded by previous teacher utterances, rather than a student utterances.

In summary, the field of education has already been significantly changed by LLMs, whose capabilities keep improving constantly. We hope that this shared task will serve to help the scientific community better understand the current capabilities of LLMs in educational contexts. Having learned from this shared task and going forward, we hope to make its future iterations even more informative.

5 Conclusion

The primary goal of this shared task was to explore the potential of the current state-of-the-art NLP and AI methods in generating teacher responses in the context of real-world teacher–student interactions. A number of diverse and strong teams participated in the task and submitted outputs of their systems to the competition, and even more people expressed their interest. The teams used a variety of the state-of-the-art large language models and explored diverse prompting and fine-tuning approaches. Importantly, these results not only shed light on the current state-of-the-art on this task but also highlighted some critical limitations that should be addressed in the future.

Acknowledgements

We thank the participants for their submissions and active involvement in this shared task. We are also grateful to them for the detailed and helpful peer reviews they provided to other shared task participants. Finally, we thank the anonymous raters on Prolific for having taken the time to provide us with additional feedback.

References

Adaeze Adigwe and Zheng Yuan. 2023. The ADAIO System at the BEA-2023 Shared Task: Shared Task 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.

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*, page to appear, Toronto, Canada. Association for Computational Linguistics.

Serge Bibauw, Wim Van den Noortgate, Thomas François, and Piet Desmet. 2022. Dialogue systems for language learning: A meta-analysis. *Language Learning & Technology*, 26(1).

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Kohd, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the Opportunities and Risks of Foundation Models. Technical report, Stanford University, Center for Research on Foundation Models (CRFM).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020a. Dialogue response ranking training with large-scale human feedback data. *arXiv preprint arXiv:2009.06978*.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020b. 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.

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*, page to appear, Toronto, Canada. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Thomas Huber, Christina Niklaus, and Siegfried Handschuh. 2023. Enhancing Educational Dialogues: A Reinforcement Learning Approach for Generating AI Teacher Responses. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications*, page to appear, Toronto, Canada. Association for Computational Linguistics.

H. W. Kuhn. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.

Margaret Li, Jason Weston, and Stephen Roller. 2019. ACUTE-EVAL: Improved Dialogue Evaluation with Optimized Questions and Multi-turn Comparisons.

Amin Omidvar and Aijun An. 2023. Empowering Conversational Agents using Semantic In-Context Learning. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications*, page to appear, Toronto, Canada. Association for Computational Linguistics.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Xavier Baró, Hugo Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. 2022. CodaLab Competitions: An open source platform to organize scientific challenges. *Technical report*.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers. In *Advances in Neural Information Processing Systems 34 Pre-Proceedings (NeurIPS 2021)*, pages 1–35.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2022. Is Reinforcement Learning (Not) for Natural Language Processing?: Benchmarks, Baselines, and Building Blocks for Natural Language Policy Optimization. *arXiv preprint arXiv:2210.01241*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Sebastian Wollny, Jan Schneider, Daniele Di Mitri, Joshua Weidlich, Marc Rittberger, and Hendrik Drachsler. 2021. Are We There Yet? - A Systematic Literature Review on Chatbots in Education. *Frontiers in Artificial Intelligence*, 4:654924.

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.

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

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.