Opportunities and Challenges in Neural Dialog Tutoring

Jakub Macina[*1,2] **Nico Daheim**[*3] **Lingzhi Wang**[4]
Tanmay Sinha[5] **Manu Kapur**[5] **Iryna Gurevych**[3] **Mrinmaya Sachan**[1]

[1]Department of Computer Science, ETH Zürich [2] ETH AI Center
[3] Ubiquitous Knowledge Processing Lab (UKP Lab), Department of Computer Science and Hessian Center for AI (hessian.AI), TU Darmstadt
[4] The Chinese University of Hong Kong
[5]Professorship for Learning Sciences and Higher Education, ETH Zürich
jakub.macina@ai.ethz.ch

Abstract

Designing dialog tutors has been challenging as it involves modeling the diverse and complex pedagogical strategies employed by human tutors. Although there have been significant recent advances in neural conversational systems using large language models (LLMs) and growth in available dialog corpora, dialog tutoring has largely remained unaffected by these advances. In this paper, we rigorously analyze various generative language models on two dialog tutoring datasets for language learning using automatic and human evaluations to understand the new opportunities brought by these advances as well as the challenges we must overcome to build models that would be usable in real educational settings. We find that although current approaches can model tutoring in constrained learning scenarios when the number of concepts to be taught and possible teacher strategies are small, they perform poorly in less constrained scenarios. Our human quality evaluation shows that both models and ground-truth annotations exhibit low performance in terms of equitable tutoring, which measures learning opportunities for students and how engaging the dialog is. To understand the behavior of our models in a real tutoring setting, we conduct a user study using expert annotators and find a significantly large number of model reasoning errors in 45% of conversations. Finally, we connect our findings to outline future work.

https://github.com/eth-nlped/dialog-tutoring

1 Introduction

The goal of dialog tutoring research is to build systems that can tutor students using natural language conversation (Wollny et al., 2021). For several decades, learning scientists have been studying the features of domain-specific dialog tutoring systems that engender learning in students (Chi et al., 1994; Graesser et al., 1995; Moore et al., 2004; Litman et al., 2006; Graesser, 2016; Ruan et al., 2019) and have established strong learning gains that are even comparable to human tutoring in specific domains (Nye et al., 2014). However, these systems require extensive authoring of materials by teachers (MacLellan and Koedinger, 2020) and therefore cannot fully utilize the scalability of online learning.

Building dialog tutors is technically challenging as tutoring dialogs typically exhibit properties that are absent in other forms of dialog. Tutoring dialogs are often *long*, enabling students to be exposed to the concepts in a way that they can use them in future (Chi and Wylie, 2014), and *grounded* in the learning scenarios (Graesser et al., 2009). Finally, good dialog tutors are engaging and create opportunities to learn, providing students space to seek and provide explanations, and self-reflect (Chi and Wylie, 2014; Reiser, 2004).

The growing success of deep neural network based language generators in other dialog settings (Adiwardana et al., 2020; Roller et al., 2021) suggests new possibilities in dialog tutoring that could scale beyond domain-specific approaches. However, despite their promise, advances in neural generative models have seen little adoption in dialog tutoring.

In this paper, we contribute a comprehensive study of the applicability of neural generative models in tutoring. We formally introduce the dialog tutoring task and analyze existing tutoring datasets (§2). Then, we describe several generative and retrieval-based models for dialog tutoring (§3) and benchmark them on two open-access dialog tutoring datasets for language learning: *CIMA* (Stasaski et al., 2020, a crowdsourced role-played dataset for learning prepositional phrases in Italian) and

[*]Equal contribution.

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2357–2372
May 2-6, 2023

Teacher-Student Chatroom Corpus (TSCC) (Caines et al., 2020, a one-to-one English tutoring dataset from an online chatroom) (§5.1). We evaluate our models on various automatic metrics (§4.2) as well as two human evaluation studies: an evaluation of the quality of the generated response with respect to various measures of goodness (§6.1), as well as a more realistic user study with a learning interface (§6.2).

Overall, while we find that pretrained models improve over simpler baselines in terms of automatic metrics, our consequent human evaluation reveals several shortcomings that ought to be addressed before these models can be adopted in the real world. We find that while neural generative models can model more constrained learning settings well, they struggle when the learning goal is more open-ended. Specifically, these models are unable to understand and reason about student solutions and misconceptions, and thus, are unable to use effective pedagogical strategies.

We find that the field of dialog tutoring is significantly limited by the quantity and quality of available datasets. The available datasets are both too small and not rich enough to capture the nuances of the dialog tutoring problem. Our analysis also reveals the inadequacy of automatic evaluation metrics for capturing tutoring quality. Not only are the existing metrics unable to capture faithfulness to the learning material and the student dialog history, but they also cannot capture moves of good human tutors that allow learners the space for reflection, explanation, follow-ups, and real engagement in the process of learning.

Based on our findings, we end with an outline of potential avenues of future research (§7). We hope that our paper will bring attention to this underexplored natural language processing application with the potential for significant social good.

2 The Dialog Tutoring Task

Dialog tutoring can be described as a multi-turn interaction between two interlocutors, where one performs the role of a *teacher* seeking to teach the other interlocutor who acts in the role of a *student*. We then can describe a dialog tutoring session formally as a sequence of turns $\mathcal{H} = (u_1, \ldots, u_{|\mathcal{H}|})$ that are taken by either of the interlocutors. Each turn $u_t \in \mathcal{V}^*$ is a finite sequence of tokens from a vocabulary $\mathcal{V}$.

Further, each turn u_t can be associated with a sequence of dialog acts $\mathbf{a}_t \in \mathcal{A}$ that indicate the action taken by the interlocutor in the corresponding turn. The dialog act is a key aspect of dialog tutoring as it can refer to the teaching strategy employed by the tutor. These may include strategies such as *providing a hint* or *seeking a clarification* (see Appendix A for more details). The set of dialog acts $\mathcal{A}$ is usually fixed according to a predefined taxonomy and may be split into two subsets $\mathcal{A} = \mathcal{A}_{\text{teacher}} \cup \mathcal{A}_{\text{student}}$, each corresponding to the teacher and student role. Each dialog session $\mathcal{H}$ may also be accompanied with some *grounding* information K, which grounds the response in relevant information and may refer to the teaching material that needs to be taught to the student. This information K may come in various formats, including images and videos. However, we restrict ourselves to using only text-based grounding in this work such that $K \in \mathcal{K} \subseteq \mathcal{V}^*$ is again a sequence of tokens from the common vocabulary $\mathcal{V}$ and $\mathcal{K}$ is used to describe the set of possible groundings (e.g., a textbook with a set of chapters). In Section 3 we derive different methods to model the role of the teacher, to which we restrict this work.

CIMA (Stasaski et al., 2020)	*TSCC (Caines et al., 2020)*
 K: "blue" is "blu" [...] Grammar Rules: Adjectives (such as color words) follow the noun they modify in Italian [...]	N/A
Teacher: (N/A) "Blue" is "blu" in Italian. Student: But what are the other words? Teacher: (N/A) Can you give me your best guess? Student: es en front de blu tree. Teacher: *(Correction)* Getting there. Remember that the adjective always follows the noun in modifies.	Teacher: *(eliciting)* So in fact fractions (half/third/quarter etc) are good to use for variety in language OK? and what about e.g. 23%? Student: just less than a quarter Teacher: *(eliciting)* so if you say 'less' you need to say 'less than' ....so just use one word ok? beginning with 'u'! Student: I am not sure of the word. Teacher: (scaffolding) just under a quarter

Figure 1: Examples of tutoring conversations from both datasets. The (image) grounding is shown in the second row and dialog acts in brackets indicate the pedagogical strategy.

2.1 Existing tutoring datasets

To our knowledge, only three conversational tutoring datasets are openly available: *CIMA* (Stasaski et al., 2020) is a crowd-sourced dataset, where annotators were asked to role-play students and teachers by working through an exercise on translating a prepositional phrase from English to Italian, given an image and a shared set of concepts. *TSCC* (Caines et al., 2020) uses real teachers leading one-on-one language tutoring sessions in English language learning, thus creating a more open-ended scenario. Finally, *TalkMoves* (Suresh et al., 2022a).

is a collection of scraped classroom transcripts of K-12 mathematics lesson videos that contain challenging, multi-party interactions.

The scarcity of tutoring datasets stands in contrast to other dialog scenarios, where plenty of datasets have been proposed and studied. For example, task-oriented dialog has been studied in domains like reservations (Wen et al., 2017; Budzianowski et al., 2018; Kim et al., 2020) or public service information (Feng et al., 2020). On the other hand, chit-chat or open-domain dialog has been studied on movies (Zhou et al., 2018), Wikipedia knowledge (Dinan et al., 2019), agent persona (Dinan et al., 2020), knowledge graphs (Moon et al., 2019), and open-ended settings (Komeili et al., 2022).

Furthermore, we note the following limitations and characteristics of tutoring datasets, also in comparison to other dialog domains: 1) Low pedagogical quality (CIMA), 2) Limited teaching strategies (all), 3) Exclusive focus on classroom settings (TalkMoves), 4) Small dataset size (all). 5) Significantly larger context sizes (TSCC) 6) Harder readability according to the Flesch score (TSCC). We provide more evidence in Table 1 which shows a comparison of dialog tutoring datasets with widely-used task-oriented and open-domain datasets.

2.2 Related work on generative dialog models

Similarly, while the advent of large pretrained models has sparked ample research on generative models for dialog (Bao et al., 2021; Peng et al., 2021; Roller et al., 2021; Shuster et al., 2022; Cohen et al., 2022), this has not carried over to research on tutoring systems, where existing solutions are predominantly rule-based and do not generate open-ended responses. For example, the authors on CIMA define heuristics to select responses (Stasaski et al., 2020). Pretrained transformers in general have only very recently been studied in this setting, however only for dialog act classification (Suresh et al., 2022b) and to study the pedagogical ability of existing large pretrained models (Tack and Piech, 2022).

3 Dialog Tutoring Models

After introducing the dialog tutoring task, this section highlights the models we evaluate on the task. We note that our aim is an analysis of existing models.

We explore turn-level models that can generate a teacher response $\mathbf{y} := u_{t+1}$ given a tutoring session $\mathcal{H} = (u_1, \ldots, u_{|\mathcal{H}|})$. During training, we obtain the dialog history by teacher forcing, i.e., we take the ground-truth dialog history. Furthermore, we do not model the problem of retrieving grounding information but rather assume it as given.

Generative Model In order to study if generative models can capture a *given* teaching strategy, we first derive a model that assumes the ground-truth dialog act sequence $\mathbf{a} = \{\mathbf{a}_1, \ldots, \mathbf{a}_{|\mathcal{H}|}\}$ to be given as an input. Then, given dialog history $\mathcal{H}_{<t} = \{u_1, \ldots, u_t\}$, grounding information K and $\mathbf{a}_{t+1} \subseteq \mathcal{A}_{\text{teacher}}$, the set of dialog acts relevant at timestep $t+1$, the teacher response $\mathbf{y}$ is generated according to a locally-normalized language generation model. In the case that no grounding information K is given, the dependency on K may be dropped.

$$\begin{aligned} \mathbf{y}^\star &= \underset{\mathbf{y}\in\mathcal{V}^*}{\operatorname{argmax}}\{p\left(\mathbf{y} \mid \mathbf{a}_{t+1}, \mathcal{H}_{<t}, K\right)\} \qquad (1) \\ &= \underset{\mathbf{y}\in\mathcal{V}^*}{\operatorname{argmax}} \prod_{i=1}^{|\mathbf{y}|}\{p\left(y_i \mid \mathbf{y}_{<i}, \mathbf{a}_{t+1}, \mathcal{H}_{<t}, K\right)\} \end{aligned}$$

We separate the turns in the dialog by special ⟨teacher⟩ and ⟨student⟩ tags and prepend the dialog act as a special token, followed by a special ⟨knowledge⟩ token and the grounding information K as the input to the encoder. In CIMA we encode the triples defining the grounding information in a simple natural language format, where we separate the English and Italian words for an object, color, and preposition as well as the whole phrase by the word "is", for example as "blue is blu" in Figure 1. Further, we add the grammar rules separated by a special token. We study different models to parametrize p that are described in Section 4.

Finally, we use the version of **CTRL** (Keskar et al., 2019) presented by Rashkin et al. (2021). The aim of the model is to improve the faithfulness of grounded response generation models, a significant problem in neural language generation (Roller et al., 2021) which holds high importance in the field of tutoring, where one trusts a teacher to present correct information. The model is augmented by a sequence of control tokens that are intended to steer the generations to desirable properties. We use the *lexical overlap* and *entailment tokens*, which we obtain as follows. In training, the lexical overlap is measured on a token-level between ground-truth response and grounding. Then, three equally sized buckets are created indicating

Dataset	Train samples	#DA	Tgt. length	Src. length	#prev. turns	corpus-div.	Flesch score	F1($\hat{\mathbf{y}}, K$)
CIMA	2,715	5	14.71	9.70	4.55	0.149	84.64	0.196
TSCC	5,845	23	16.09	11.72	68.28	0.327	73.00	-
MultiWoZ 2.1	56,781	34*	19.86	14.49	7.86	0.069	90.90	-
Schema-Guided Dialog	164,982	10*	14.30	10.36	11.38	0.049	95.37	-
DSTC9	19,184	-	21.61	11.65	11.70	0.050	81.85	0.47
Personachat	127,162	-	12.26	11.65	6.51	0.162	91.80	0.10
FaithDial	18,357	-	21.72	17.33	4.54	0.274	83.28	0.47
CMU_DoG	81,468	-	14.49	18.23	18.73	0.178	79.54	0.02

Table 1: Statistics of dialogue datasets with lines separating groups by task - tutoring, task-oriented, open-domain dialog tasks. Target length and source length in average number of tokens (Bart tokenizer), # prev. turns is averaged for each teacher response, corpus-div is ngram entropy averaged for uni to four-grams. * We only count system dialog acts.

low, medium, and high overlap which is indicated by a control token. Entailment is determined by an MNLI model and again a corresponding token is added. At test time, we always use the token that encourages the desirable property, in this case high lexical overlap and entailment. Finally, using a sequence of control tokens $\mathbf{c}$, the model from equation 1 becomes:

$$p\left(\mathbf{y} \mid \mathbf{a}_{t+1}, \mathbf{c}, \mathcal{H}_{<t}, K\right) \tag{2}$$

Joint Model In order to study how well current neural models can decide on a reasonable teaching strategy and perform in real case scenarios, we also implement a model that first decides the dialog act $a_{t+1} \in \mathcal{A}_{\text{teacher}}$ (instead of assuming the ground-truth dialog act) and then uses it to generate a response $\mathbf{y} = u_{t+1}$. We use a simple model that again takes the grounding and dialog context as input but now generates the concatenation of dialog act and response in one utterance, akin to SOLOIST (Peng et al., 2021). Thus, for a given $\tilde{\mathbf{y}} := \mathbf{a}_{t+1} \circ \mathbf{y}$ with act sequence $\mathbf{a}_{t+1}$ of length N and response $\mathbf{y}$ of length T, the model is

$$p\left(\tilde{\mathbf{y}} \mid K, \mathcal{H}_{<t}\right) = \prod_{i=1}^{m+N} p\left(\tilde{y}_i \mid \tilde{\mathbf{y}}_{<i}, K, \mathcal{H}_{<t}\right) \tag{3}$$

In training, we use teacher forcing and prepend $\mathbf{a}_{t+1}$ to $\mathbf{y}$ to obtain the label sequence. At test time, the model performs a beam search over the dialog act sequence and response jointly.

Retrieval-based model Since generative models are known to produce erroneous outputs that are factually incorrect and potentially inappropriate (Ji et al., 2022), we also experiment with using a retrieval-based model that selects responses from the training corpus at test time. As opposed to previous work on the topic (e.g., Stasaski et al. (2020)), we do not employ a rule-based model but rather a learned retrieval model that does not require handcrafting elaborate and possibly brittle rules. Therefore, we use the **Bi-Encoder** architecture (Mazaré et al., 2018; Dinan et al., 2019) where a dialog context encoder $\text{enc}_{\mathcal{H}_{<t};\theta}$ and a response encoder $\text{enc}_{\mathbf{y};\theta}$ encode context $\mathcal{H}_{<t}$ and possible responses $\mathbf{y}$ into a fixed size vector of same dimension n. In our experiments, the weights θ of both encoders are shared.

The model is trained using contrastive learning. Suppose we are given a training pair $\mathcal{H}, \hat{\mathbf{y}}$ from a training dataset $\mathcal{D}$ that we use for teacher forcing. We then train the model by sampling a negative response $\bar{\mathbf{y}}$ from the set of responses in $\mathcal{D}$ and using the Triplet Loss criterion, which for a metric function $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is defined as:

$$\begin{aligned}\mathcal{L}(\theta; \mathcal{H}, \hat{\mathbf{y}}, \bar{\mathbf{y}}) = [m + d(\text{enc}_{\mathcal{H};\theta}(\mathcal{H}), \text{enc}_{\mathbf{y};\theta}(\hat{\mathbf{y}})) \\ - d(\text{enc}_{\mathcal{H};\theta}(\mathcal{H}), \text{enc}_{\mathbf{y};\theta}(\bar{\mathbf{y}}))]_+,\end{aligned} \tag{4}$$

where m is a margin hyperparameter, and d is the euclidean norm in our experiments. Further, we do stratified sampling on CIMA to not select negatives from the same preposition, color, or object that might be false negatives. At test time, given a dialog context $\mathcal{H}_{<t}$, we choose a response $\mathbf{y}^\star$ from the training set $\mathcal{D}$ by maximum inner product search using the decision rule

$$\mathbf{y}^\star = \underset{\mathbf{y} \in \mathcal{D}}{\text{argmax}} \{\text{enc}_{\mathcal{H};\theta}(\mathcal{H}_{<t})^T \text{enc}_{\mathbf{y};\theta}(\mathbf{y})\}. \tag{5}$$

4 Experiments

We use the following models for parameterizing p in Equation 2: A **sequence-to-sequence** model (Sutskever et al., 2014) with a copy mechanism (Gu et al., 2016) trained from scratch. A wide range of pretrained Transformers, namely **BART** (Lewis et al., 2020), **DialoGPT** (Zhang et al., 2020), **T5** (Raffel et al., 2020) and its multilingual version **mT5** (Xue et al., 2021).

Model	CIMA			TSCC	
	sBLEU / BLEU-1 (↑)	BERT F1 (↑)	Q^2 (↑)	sBLEU / BLEU-1 (↑)	BERT F1 (↑)
Rule-based (Stasaski et al., 2020)*	0.34/-	0.45	-	-	-
LSTM (Stasaski et al., 2020)*	0.31/-	0.53	-	-	-
Seq2seq	2.89 / 28.0	0.676	0.372	-	-
DialoGPT	4.12 / 35.6	0.697	0.571	0.63 / 18.5	0.661
Bi-Encoder (RoBERTa-base)	5.89 / 23.9	0.690	0.344	1.367 / 8.8	0.638
CTRL (BART-base)	5.99 / 42.5	0.702	0.673	-	-
t5-small	7.36 / 34.0	0.672	0.676	2.72 / 12.1	0.646
BART-large	8.61 / 38.7	0.715	0.673	1.85 / 13.7	0.658
BART-base	9.58 / **42.5**	**0.726**	**0.680**	**2.67 / 18.6**	**0.670**
mt5-small	**11.26** / 41.0	0.700	0.624	1.80 / 14.9	0.653
BART-base†	5.61 / 41.03	0.707	0.642	1.90 / 15.4	0.659
BART-large†	5.65 / 42.67	0.694	0.607	1.74 / 15.1	0.660

Table 2: Comparison of models on CIMA and TSCC. We note that the strong sacrebleu differences are caused by the brevity penalty (all generative models generate too short sequences), †: use predicted dialog act label, others use ground-truth. * numbers taken from (Stasaski et al., 2020) which may not be comparable as there is no standard split of CIMA dataset.

BART and T5 are pretrained encoder-decoder models that were trained on denoising and text-to-text tasks, respectively. mT5 bases on T5 but is multilingual which might help with the code-switched utterances in CIMA. Lastly, DialoGPT is an autoregressive language model based on GPT-2 (Radford et al., 2019) that was pretrained on a large dialog dataset obtained from Reddit. With this, we intend to study whether large-scale dialog-specific pretraining can aid in training educational tutors, as well.

Implementation Details We implement our experiments using the Huggingface transformers library and finetune the checkpoints provided as part of it for all Transformer-based models. For these models, we use an initial learning rate of 3.25×10^{-5}, 500 warmup steps and linear learning rate decay. We train the models using a batch size of 8 and evaluate on the validation sets after each epoch. In the end, we select the best model to be the one that has a minimal loss on the validation set. The sequence-to-sequence baseline is trained from scratch using an initial learning rate of 0.001 for 25,000 steps using the Adam optimizer and a dropout rate of 0.1 We use beam search with a beam size of 10 to generate model responses.

4.1 Dataset splits

Since there are no official dataset splits for CIMA and TSCC, we split both datasets randomly into training, validation and test sets. We provide the exact split of the dataset in an accompanying code repository. For CIMA, we use all such samples with less than three annotated tutor responses for training. The other conversations are split randomly into equally-sized validation and test sets which results in 2715/300/300 samples each.

For TSCC, we split randomly along the conversations to obtain 82/10/11 training, validation, and test conversations each.

4.2 Evaluation metrics

To evaluate our models, we use the BLEU implementation provided by the sacrebleu package (sBLEU) (Post, 2018) to measure lexical overlap between generated and ground-truth response. Furthermore, we use BERT F1 (BERTScore) to measure their semantic similarity. Lastly, for CIMA we also calculate Q^2 (Honovich et al., 2021) which measures the factual consistency of the response $\mathbf{y}$ with the grounding information K by employing a question-answering based matching. Both BERTScore and Q^2 have shown strong correlation with human judgements on factual consistency Honovich et al. (2022).

5 Results

In this section, we summarize our main findings in terms of automatic evaluation. First, we give an overview of the performance of different models that we train on CIMA and TSCC in Section 5.1. Then, we assess their ability to stay faithful to teaching strategies (Section 5.2) and study how grounding annotations can influence the faithfulness of neural dialog tutors (Section 5.3), before studying their scaling behavior with dataset size and complexity (Section 5.4) and their generalization capabilities (Section 5.5). We then finish with an assessment of using education-specific data for pretraining (Section 5.6).

	Method				
	GT	$\text{BART}_{\text{base}}$	$\text{BART}_{\text{large}}$	CTRL	Retrieval
DA F1	78.3	**81.0**	70.1	63.0	43.1

Table 3: F1 score of the dialog act classification based on the generated responses of our models.

5.1 Comparison of different models

Table 2 shows the key results from our experiments. First, all automatic metrics are *significantly higher* on CIMA, which indicates that the models can fit CIMA much better than TSCC, with which current approaches still struggle. We further analyse this finding in Section 5.2 and show that this is because TSCC has *richer teaching strategies which are harder to model*. Our comparison also suggests that finetuning large pretrained *Transformer models generally gives better results than the rule-based and LSTM model* reported in (Stasaski et al., 2020), and our implemented retrieval and sequence-to-sequence baselines. This illustrates the potential of LLMs for dialog tutoring.

We also see a significant difference among different LLMs. Dialog-specific pretraining of DialoGPT does not help and gives worse results than BART and T5, primarily because the model tends to generate short and generic responses more often. Multilingual pretraining in mT5 improves over T5 only in some metrics, notably in BLEU and BERT F1 on CIMA but not in terms of Q^2. Similarly, adding control tokens to BART does not improve Q^2 or other automatic metrics. Surprisingly, using very large models actually degrades performance in our experiments. Finally, the last two rows show results obtained with our joint model that does not use the ground-truth dialog act but predicts it together with the response sequence and still provides reasonable performance.

5.2 How well can generative models capture teaching strategies?

We study this question first by evaluating the dialog act prediction accuracy of our joint model. We find that it is *significantly low* on TSCC with 21.8 compared to 71.2 on CIMA for BART-base which indicates significant room for improvement. Notably, the joint model tends to predict more frequently occuring dialog acts, which results in fewer follow-up questions and "Other" never being predicted in CIMA, the least frequent act in the data. The distribution of dialog acts in the ground-truth annotations and model predictions with a BART-base joint model is in Figure 2.

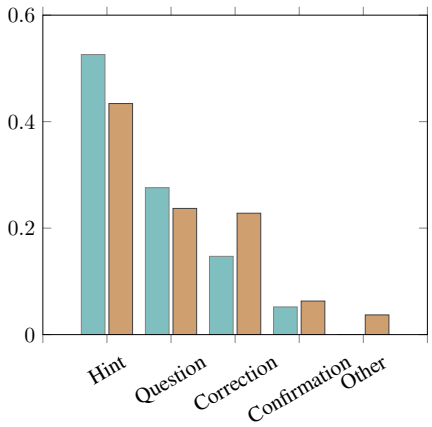


Figure 2: Distribution of predicted and ground-truth dialog acts on CIMA.

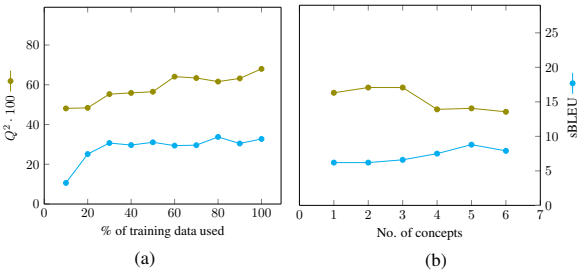
Model	sBLEU (↑)	BERT F1 (↑)	Q^2 (↑)
BART-base	6.69 / 38.6	0.718	0.571
+ triples	9.20 / **45.3**	**0.730**	0.642
+ grammar rules	**9.58** / 42.5	0.726	**0.680**

Table 4: Comparison of models with different inputs on CIMA. Triples are made up of preposition, object, and color translations. Grammar rules are a textual description of a learning concept.

Then, we evaluate how well different models can stick to a given ground-truth dialog act by predicting the dialog act of the *generated response* with a BART-base model trained to predict the ground-truth dialog act sequence based on the ground-truth response. The results are shown in Table 3. Notably, *BART-base performs better than the ground-truth annotations*. The CTRL model, on the other hand, has worse performance since the control tokens do not respect tutoring principles (e.g., lexical overlap to grounding discourages follow-up questions in favor of just giving hints).

5.3 Does grounding in learning concepts help?

Prior work has shown that grounding responses in relevant data can improve their quality, especially in terms of faithfulness (Shuster et al., 2021). We intend to validate this for dialog tutoring by studying three models with different inputs on CIMA. The first model is not provided grounding information, whereas the second and third are grounded in learning concepts (cf. Equation 1) with one using only the (preposition, object, color) triples and the other making use of additional grammar rules. The results with these models are shown in Table 4 and suggest that *grounding responses in relevant knowledge helps the model to produce better and*



(a) (b)

Figure 3: Performance of BART-base on CIMA as a function of: (a) training data size uniformly sampled from the training data, (b) the number of concepts, where only the specific number of concepts is retained and all others are excluded.

more faithful responses.

5.4 How do models scale with more data?

Due to the limited availability of high-quality pedagogical datasets and the time-consuming process of authoring new materials (MacLellan and Koedinger, 2020), it is important to understand how quickly generative models can generalize to new settings. Thus, we assess how well the model can model tutoring in low-resource scenarios. We construct a study, where we randomly sample subsets of the CIMA training set and test the performance of the various models. We can see from Figure 3(a) that with more training data, the faithfulness of responses appears to improve and is not saturated before we reach the full training set. This supports the intuition that *additional training data might improve the performance further*.

Similarly, we study how well our model can deal with an increase in complexity with respect to learning concepts at similar training data sizes. Therefore, we construct different training datasets with 735 samples and a varying number of concepts each time. We begin by taking samples concerned with the concept "in front of the" and evaluate exclusively on it, gradually adding new concepts. Figure 3(b) suggests that Q^2 drops sharply at four concepts. BLEU on the other hand increases, and this might be due to the metric encouraging generic utterances that, for example, repeat a grammar rule.

5.5 Can models generalize to new concepts?

As the students progress and gain new knowledge, it might be a desirable property of dialog tutoring models to be able to handle new concepts that suit this increase in prior knowledge. Hence, we study how well our CIMA model can generalize to new concepts that it has not seen in training, for example, a new preposition. For this analysis, we create a set-up where we first train the model on all of the training data and evaluate on the subset of samples for each preposition separately. We then compare this number to a model that is not trained on the corresponding concept it is evaluated on, creating a zero-shot set-up which we carry out for a grounded and ungrounded response generation model. As measured by Q^2 (cf. Table 5), this model can indeed *generalize to new concepts well, albeit with performance degradation*. Furthermore, *grounding information improves generalization* as these define the learning concept (in this case the preposition) and how it is used. Without this information, we observe that the model generates generic responses more often.

Concept	#Samples	Full data	Zero-shot	Zero-shot without grounding
	train/test	Q^2	Q^2	Q^2
is behind the	549/90	0.698	0.603	0.533
is in front of the	735/84	0.616	0.512	0.500
is next to the	547/51	0.497	0.539	0.483
is on top of the	224/30	0.683	0.578	0.567
is under the	270/24	0.854	0.646	0.625
is inside of the	390/21	0.579	0.643	0.190
all concepts	2715 / 300	0.644	0.570	0.502

Table 5: Performance of a grounded BART-base model by learning concept. Full data uses the entire training data and zero-shot removes the concept of the row from the training data.

Method	sBLEU	BERT F1	Q^2
BART-base	6.69 / 38.6	0.718	0.571
+ Ed. data	**7.31 / 41.4**	**0.727**	0.577
+ Non-Ed. data	6.60 / 39.4	0.721	**0.583**

Table 6: Influence of pretraining on educational and non-educational data. Please note that no grounding information is used in this setting.

5.6 Does education-specific pre-training help?

As educational data are widely available on the internet, next we study how education-specific pre-training effect results. In Table 6, we show results obtained with finetuning a BART-base model directly on CIMA and pretraining it on tutoring dialogs from TSCC or non-tutoring dialogs from MultiWoZ 2.1 (Eric et al., 2020), Personachat (Zhang et al., 2018), CMU DoG (Zhou et al., 2018), DSTC9 (Kim et al., 2020) and Topicalchat (Gopalakrishnan et al., 2019). In both cases, we only see *minor improvements*, which may be explained by the different dataset settings and the lack of a unified dialog act taxonomy.

6 Human Evaluation

We further evaluate previously assessed models with human judgments firstly by obtaining quality estimates according to different criteria and secondly by conducting a simulation study, where expert annotators are asked to provide novel rewritings of existing conversations and to categorize errors made by the model.

6.1 Quality of the generated responses

We perform a human quality evaluation of the generated response for four models - retrieval (Bi-Encoder), BART-base, BART-base$_{\text{CTRL}}$ and the joint model (BART-base). A randomly chosen subset of the CIMA test set conversations were annotated by 4 annotators (with one annotator speaking C1 level Italian). All annotators labeled 60 examples in total, of which 20 overlapped. To further distinguish the quality of training data for the models, we annotated ground-truth responses on a small sample of 20 examples. We evaluate the following criteria on a 3-point Likert scale (disagree to completely agree) and outline our findings in the following, as shown in Figure 4.

Fluency *"The response is grammatically correct and fluent."* We find that all models have very high fluency scores.

Coherence *"The response naturally follows up on previous utterance and context and has no logical conflicts with the context or DA label."* We find that all generative models are able to produce coherent responses but not the retrieval model.

Correctness *"The response is factually correct and respects learning concepts being taught."* All models score comparable to ground-truth responses on the constrained CIMA dataset. It is noteworthy, however, that a response may be correct in itself but not coherent with the context or the grounding (often the case in the retrieval model), and this could explain the discrepancy between correctness and our automatic Q^2 scores.

Equitable tutoring *"The response gives a learning opportunity for the student by providing space for reflection, explanation, pointing to follow-up challenge, or engaging student in other ways."* Here we find significant deficiencies not only for our evaluated models but notably also for the annotated ground-truth responses (gt). This might explain the insufficiencies in the responses as they reflect this distributional behavior of the training data. We think that future dataset collections should take better care of this property and resort to more expert annotators as opposed to crowdsourcing.

Quality Attribute	sBLEU	BERTScore
Fluency	0.14	0.12
Coherence	0.17	0.26
Correctness	0.06	0.15
Equitable Tutoring	0.08	0.16

Table 7: Pearson correlation coefficients between the human judgements on our quality criteria and automatic metrics.

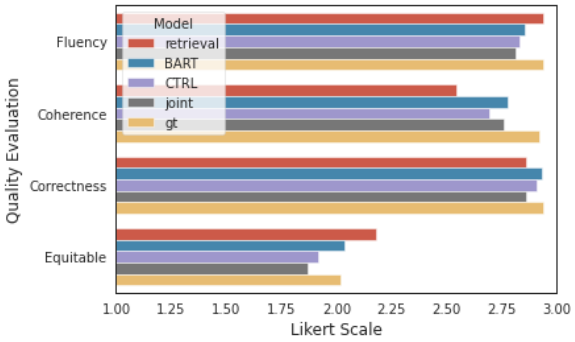


Figure 4: Comparison of models on four criteria (reporting M) in the human quality evaluation. We observed high SD for coherence and equitable metrics.

Furthermore, Table 7 shows that our automatic metrics correlate poorly with human judgements.

6.2 User study with a learning interface

Lastly, we seek to study how well dialog tutoring models can perform in a realistic setting with questions obtained from real users (containing out-of-distribution samples) and not the fixed dataset. Therefore, we randomly sampled conversations from the CIMA test set. We asked two C1-level expert Italian speakers to 1) rephrase these conversations using a conversational dialogue interface and 2) assign erroneous model responses to predefined error categories. The interface used in the qualitative evaluation is shown in Figure 6. We obtain all model responses from the BART-base model that first predicts the dialog act and then the response. Error categories adopted from previous work (Bommasani et al., 2021) describe the ideal behavior of tutoring models as simulating the behavior of good human teachers along two dimensions:

Understanding *"Being able to understand and reason about student solutions, misconceptions, and learning concepts."* We find that of the 20

modified conversations, 45% exhibit *Understanding errors*, such as an incorrect solution assessment or incorrect translations.

Pedagogy *"Being able to use effective pedagogy to instruct students."* We find that 10% of the responses exhibit *Pedagogical errors*, for example telling the correct solution directly without offering any engagement point to the student.

50% of the conversations were labeled good by the annotators. Examples of the conversations are available in Table 8.

7 Discussion: Towards More Equitable and Faithful Tutoring Systems

In this section, we outline directions of research that we think can be important steps towards more equitable and faithful tutoring models. Namely, we first address the small scale and quality of current tutoring datasets and cast doubt on the crowdsourcing data quality checks. Then, we suggest ways of improving the underperformance of both equitable tutoring and teaching strategy prediction identified in current generative models under these constraints by drawing from learning sciences literature. Finally, we outline desiderata for more reliable dialog evaluation of neural tutoring models.

Datasets Based on the analysis in §2.1 and Table 1, we think that the community will benefit from a dataset that lies between CIMA and TSCC in terms of its difficulty. Moreover, the low equitable tutoring scores of CIMA's ground-truth responses indicate that crowdsourcing with untrained annotators can lead to low pedagogical quality. A similar observation has been found by human evaluation for the TSCC dataset (Tack and Piech, 2022). Finally, we encourage the establishment of better dialog act taxonomies that are backed by learning sciences research. As outlined in §5.6 and in He et al. (2022), a unified taxonomy may also strongly aid in transfer learning.

Models So far, dialog tutoring models have only covered limited domain-specific settings linked to a particular activity, such as learning Italian prepositions or solving math word problems. We argue that the community could benefit from working on problems common to learning in general, for example tracking problem-solving states and modeling pedagogies used by teachers. Here, knowledge tracing (Corbett and Anderson, 1994) (the problem of estimating students' skill mastery level) could be used for tracking problem-solving states and increasing the coherence of dialog tutoring conversations and dialog act selection performance which would contribute to better modeling of global teaching strategies. Furthermore, validated instruction quality coding schemes (Michaels et al., 2010; Hennessy et al., 2016) used by classroom teachers can be computationally modeled (Demszky et al., 2021; Ganesh et al., 2021) and incorporated into models.

We also think that recently proposed constrained decoding approaches that can balance between multiple criteria (Qin et al., 2022) hold great promise in improving faithfulness in complex tutoring dialogs. Finally, as data collection is labor-intensive in expert domains, we see great potential in few-shot learning methods, such as prompt-based methods (Schick and Schütze, 2022).

Evaluation Our experiments highlight the insufficiencies of current automatic dialog evaluation metrics, as both BLEU and BertScore show comparatively low correlation with our collected human judgements from §6.1. This is in line with previous research (Mehri and Eskenazi, 2020; Mehri et al., 2022) and shows the necessity not only for better automatic evaluation metrics but also for verification based on human judgements or user studies that should incorporate criteria relevant to tutoring (e.g., equitable tutoring outcomes). Metrics that incorporate task success, which have been used in task-oriented dialog systems (Budzianowski et al., 2018), are a promising direction of future research for automatic evaluation.

8 Conclusion

In this work, we reflected on the state of research in dialog tutoring and explored the potential of neural generative models in this domain. We found some promising initial results with these models in comparison to rule- or retrieval-based methods. However, we also established limitations of currently available benchmarks and evaluation criteria. Furthermore, we showed that there are a number of challenges that need to be addressed before neural generative models of text can be deployed as intelligent tutoring systems on a larger scale, such as controllability and being able to model a sound pedagogical strategy. Based on these findings, we outline potential avenues for future research.

Limitations

A key limitation of our work is the use of only two available tutoring datasets. Despite a limited number of datasets available in this domain, using the TalkMoves dataset (Suresh et al., 2022a) could help further generalize our findings. This remains an avenue for future work.

Based on the prior work, we focused on the specific conversational goal of dialog tutors which is providing learning aid for students' skill development and more opportunities to learn. While this is the most widespread type (Wollny et al., 2021), it is not covering all the goals of human tutors, and other aspects could be important, for example, rapport-building or mentoring on the meta-cognitive level. We acknowledge this both as a prerequisite for our work and at the same time as a limitation. For further discussion we refer the reader to Appendix B and C.

Finally, our user study could be further extended with more participants. In the future, we plan a more comprehensive study with real language learners using an end-to-end dialog tutoring system.

Ethics Statement

We do not foresee any significant harm directly as a result of our work. Having said that, we must understand that automatic tutoring is a high-stake setting that can pose significant harm if appropriate care is not taken before the deployment of these systems. Issues of biases and lack of trust, and other ethical issues such as privacy concerns must be considered. Considering learners only as data points within a neural dialog tutoring context may prevent us from seeing the societal and socioeconomic barriers that they may be up against, thereby running the risk of not only failing to help relevant learner subgroups but also sometimes giving additional privileges to those who use these systems.

9 Acknowledgements

This project was made possible by an ETH AI Center Doctoral Fellowship to Jakub Macina with partial support by the Asuera Stiftung and the ETH Zurich Foundation and has received funding by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE. We thank the group members and our reviewers for their valuable feedback.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *ArXiv preprint*, abs/2001.09977.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2021. PLATO-2: Towards building an open-domain chatbot via curriculum learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2513–2525, Online. Association for Computational Linguistics.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

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

Andrew Caines, Helen Yannakoudakis, 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.

Michelene TH Chi, Nicholas De Leeuw, Mei-Hung Chiu, and Christian LaVancher. 1994. Eliciting self-explanations improves understanding. *Cognitive science*, 18(3):439–477.

Michelene TH Chi and Ruth Wylie. 2014. The icap framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, 49(4):219–243.

Aaron Daniel Cohen, Adam Roberts, Alejandra Molina, Alena Butryna, Alicia Jin, Apoorv Kulshreshtha, Ben Hutchinson, Ben Zevenbergen, Blaise Hilary Aguera-Arcas, Chung ching Chang, Claire Cui, Cosmo Du, Daniel De Freitas Adiwardana, Dehao Chen, Dmitry (Dima) Lepikhin, Ed H. Chi, Erin Hoffman-John, Heng-Tze Cheng, Hongrae Lee, Igor Krivokon, James Qin, Jamie Hall, Joe Fenton, Johnny Soraker, Kathy Meier-Hellstern, Kristen Olson, Lora Mois Aroyo, Maarten Paul Bosma,

Marc Joseph Pickett, Marcelo Amorim Menegali, Marian Croak, Mark Díaz, Matthew Lamm, Maxim Krikun, Meredith Ringel Morris, Noam Shazeer, Quoc V. Le, Rachel Bernstein, Ravi Rajakumar, Ray Kurzweil, Romal Thoppilan, Steven Zheng, Taylor Bos, Toju Duke, Tulsee Doshi, Vinodkumar Prabhakaran, Will Rusch, YaGuang Li, Yanping Huang, Yanqi Zhou, Yuanzhong Xu, and Zhifeng Chen. 2022. Lamda: Language models for dialog applications. In *arXiv*.

Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4:253–278.

Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori Hashimoto. 2021. Measuring conversational uptake: A case study on student-teacher interactions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1638–1653, Online. Association for Computational Linguistics.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition*, pages 187–208. Springer.

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

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online. Association for Computational Linguistics.

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

Ananya Ganesh, Martha Palmer, and Katharina Kann. 2021. What would a teacher do? Predicting future talk moves. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4739–4751, Online. Association for Computational Linguistics.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Arthur C Graesser. 2016. Conversations with autotutor help students learn. *International Journal of Artificial Intelligence in Education*, 26(1):124–132.

Arthur C Graesser, SIDNEY D'MELLO, and Natalie Person. 2009. Meta-knowledge in tutoring. In *Handbook of metacognition in education*, pages 373–394. Routledge.

Arthur C Graesser, Natalie K Person, and Joseph P Magliano. 1995. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied cognitive psychology*, 9(6):495–522.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. 2022. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Sara Hennessy, Sylvia Rojas-Drummond, Rupert Higham, Ana María Márquez, Fiona Maine, Rosa María Ríos, Rocío García-Carrión, Omar Torreblanca, and María José Barrera. 2016. Developing a coding scheme for analysing classroom dialogue across educational contexts. *Learning, culture and social interaction*, 9:16–44.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021.

Q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *ACM Computing Surveys*.

Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL - A Conditional Transformer Language Model for Controllable Generation. *ArXiv preprint*, abs/1909.05858.

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

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.

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

Diane J Litman, Carolyn P Rosé, Kate Forbes-Riley, Kurt VanLehn, Dumisizwe Bhembe, and Scott Silliman. 2006. Spoken versus typed human and computer dialogue tutoring. *International Journal of Artificial Intelligence in Education*, 16(2):145–170.

Christopher J MacLellan and Kenneth R Koedinger. 2020. Domain-general tutor authoring with apprentice learner models. *International Journal of Artificial Intelligence in Education*, pages 1–42.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

Shikib Mehri, Jinho Choi, Luis Fernando D'Haro, Jan Deriu, Maxine Eskenazi, Milica Gasic, Kallirroi Georgila, Dilek Hakkani-Tur, Zekang Li, Verena Rieser, et al. 2022. Report from the nsf future directions workshop on automatic evaluation of dialog: Research directions and challenges. *ArXiv preprint*, abs/2203.10012.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Sarah Michaels, Mary Catherine O'Connor, Megan Williams Hall, and Lauren B Resnick. 2010. Accountable talk sourcebook: For classroom conversation that works. *Pittsburgh, PA: University of Pittsburgh Institute for Learning*.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. ParlAI: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.

Johanna D Moore, Kaska Porayska-Pomsta, Sebastian Varges, and Claus Zinn. 2004. Generating tutorial feedback with affect. In *FLAIRS Conference*, pages 923–928.

Benjamin D Nye, Arthur C Graesser, and Xiangen Hu. 2014. Autotutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, 24(4):427–469.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. Soloist: Building task bots at scale with transfer learning and machine teaching. *Transactions of the Association for Computational Linguistics*, 9:807–824.

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

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. 2022. COLD decoding: Energy-based constrained text generation with langevin dynamics. In *Advances in Neural Information Processing Systems*.

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

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

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. 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 704–718, Online. Association for Computational Linguistics.

Brian J. Reiser. 2004. Scaffolding Complex Learning: The Mechanisms of Structuring and Problematizing Student Work. *Journal of the Learning Sciences*, 13(3):273–304. Publisher: Routledge _eprint: https://doi.org/10.1207/s15327809jls1303_2.

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.

Jeremy Roschelle and Stephanie D Teasley. 1995. The construction of shared knowledge in collaborative problem solving. In *Computer supported collaborative learning*, pages 69–97. Springer.

Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. 2019. Quizbot: A dialogue-based adaptive learning system for factual knowledge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 357. ACM.

Timo Schick and Hinrich Schütze. 2022. True few-shot learning with prompts—a real-world perspective. *Transactions of the Association for Computational Linguistics*, 10:716–731.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *ArXiv preprint*, abs/2208.03188.

Tanmay Sinha and Manu Kapur. 2021. When problem solving followed by instruction works: Evidence for productive failure. *Review of Educational Research*, 91(5):761–798.

Katherine Stasaski, Kimberly Kao, and Marti A. Hearst. 2020. CIMA: A large open access dialogue dataset for tutoring. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–64, Seattle, WA, USA â†' Online. Association for Computational Linguistics.

Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H Martin, and Tamara Sumner. 2022a. The talkmoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. *arXiv preprint arXiv:2204.09652*.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

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.

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

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.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual

pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

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

Ran Zhao, Tanmay Sinha, Alan W Black, and Justine Cassell. 2016. Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In *International conference on intelligent virtual agents*, pages 218–233. Springer.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.

A Pedagogical strategy and dialog acts in dialog tutoring

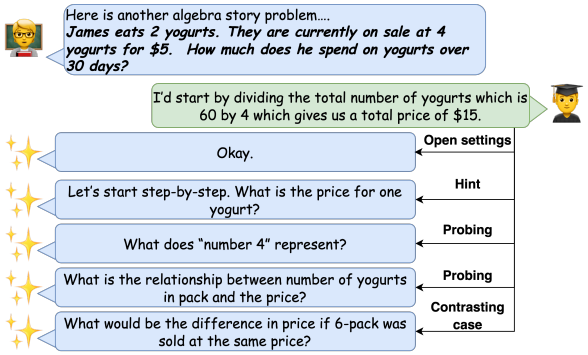


Figure 5: Example dialogue between a tutor and a student solving an algebra story problem. Key questions are: What teacher pedagogical strategies are the best in terms of learning gains of students? How to adapt language models to generate pedagogically valid responses?

In the context of this paper, we assume that the pedagogical strategy is represented using dialog act annotations. An example of the teacher strategy is providing hints (cf. example in Figure 5), where a teacher provides helpful support or clarifies goals to the student. Another example is Probing (cf. example in Figure 5), which prompts students to explain better or reflect on the current solution. CIMA contains five teacher dialog acts - *hint, open-ended question, correction, confirmation, other*. TSCC contains more fine-grained dialog acts such as *eliciting, scaffolding, enquiry, or recap*.

From a learning science standpoint, pedagogical strategy could be viewed as a global strategy (knowing how to effectively guide students e.g. using questioning or providing contrasting cases) and dialog acts as a specific decision on how this strategy is implemented on the local turn-based level.

B Equitable tutoring

Although tutoring is typically conceived as a scenario where a subject matter expert works synchronously with one or multiple students and takes interpretive authority, there is increasing empirical evidence supporting the case for incorporating active learning approaches in the classroom (Freeman et al., 2014; Sinha and Kapur, 2021). With collaborative creation of knowledge where teachers position themselves as co-learners and students also take interpretive authority, such approaches are better poised to build classroom equity than monologic educational practices where only one voice (primarily the teacher's) tends to be heard, legitimized and sometimes imposed. Therefore, if we rethink of the goals of education as providing opportunities for students to enter into the workforce with a positive identity about themselves and the subject matter, equitable tutoring via increased student chances to pose ideas, construct knowledge and as a result feel welcomed into the intellectual discussion, holds tremendous promise.

C Conversational Goals

In this work, we studied only 1:1 dialog tutoring settings with a specific focus on the role of a teacher/tutor. We focused on the most commonly used goal of dialog tutoring which is a learning aid to support students' skill development and provide opportunities to learn (Wollny et al., 2021).

However, teacher-student interactions may entail multiple conversational goals that can serve interactional functions (e.g., turn-taking) and interpersonal functions (e.g., rapport-building moves such as self-disclosure, praise, social norm violation). Research in human tutoring and collaborative learning, more generally, has shown that how students manage the task space (consisting of the problem to be solved, e.g., do I agree with the interlocutor's reasoning?) and the relational space (consisting of the interactional challenges and opportunities, e.g., can I disagree without threatening the interlocutor's face?) is critical to learning outcomes (Roschelle and Teasley, 1995; Zhao et al., 2016). Neural dialog tutoring, which can account for this crucial, but fundamental distinction among conversational goals is challenging.

D Qualitative user evaluation

The interface used in the qualitative evaluation is shown in Figure 6.

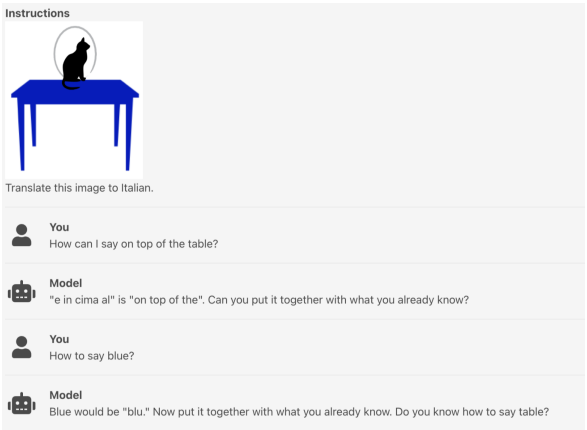


Figure 6: User interface used for the qualitative evaluation. The interface is based on ParlAI (Miller et al., 2017) library.

Context	Generated teacher response
T: Please try to fill in the blank in Italian. S: how do you say next? T: Remember that "is next to the" is "e vicino" S: I don't know how to say purple. Is it viola?	Viola is the word for purple.
T: Please try to fill in the blank in Italian. S: how do you say next? T: Remember that "is next to the" is "e vicino" S: Purple is viola, right?	That is correct.
T: Please try to fill in the blank in Italian. S: how do you say next? T: Remember that "is next to the" is "e vicino" S: I think purple could be violett.	Viola is the word for purple.
T: Please try to fill in the blank in Italian. S: how do you say next? T: Remember that "is next to the" is "e vicino" S: I can't remember the color. Can you give me a hint?	Purple is viola.
T: Please try to fill in the blank in Italian. S: how do you say next? T: Remember that "is next to the" is "e vicino" S: I guess it is e vicino all'albero. Is it correct to start the sentence with the preposition?	The preposition for "purple" is "viola" **Understanding - Meaning mismatch**
T: "Blue" is "blu". Please try to fill in the blank in Italian. S: But what are the other words? T: Can you give me your best guess? S: e di fronte all'albero blu (**correct**)	Almost, tree is all'albero'. Please try again. **Understanding - Incorrect solution assessment**
T: "Blue" is "blu". Please try to fill in the blank in Italian. S: But what are the other words? T: Can you give me your best guess? S: e di fronte blu all'albero (**adjective position swap**)	Almost, you need to add blu after "all'albero" to make the word "tree". **Understanding - Incorrect translation**
T: "Blue" is "blu". Please try to fill in the blank in Italian. S: But what are the other words? T: Can you give me your best guess? S: e di fronte al albero blu (**wrong definite article**)	Almost, tree is all'albero not albero.
T: "Blue" is "blu". Please try to fill in the blank in Italian. S: But what are the other words? T: Can you give me your best guess? S: e di fronte al cane blu (**noun swap**)	Almost. Al cane is not the correct word for tree.
T: "Blue" is "blu". Please try to fill in the blank in Italian. S: But what are the other words? T: Can you give me your best guess? S: e dietro all'albero blu" (**wrong preposition**)	The correct answer is e di fronte all'albero blu. **Pedagogy - Information reveal**

Table 8: Examples of rephrased conversations from the qualitative study. T refers to a teacher utterance, S refers to a student utterance. Bold text is information for the reader indicating error categories.