Opportunities and Challenges in Neural Dialog Tutoring
Jakub Macina∗1,2  Nico Daheim∗3  Lingzhi Wang4
Tanmay Sinha5  Manu Kapur5  Iryna Gurevych3  Mrinmaya Sachan1

1Department 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  5Professorship 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
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 under-explored natural language processing application with the potential for significant social good.

# 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 $H = (u_1, \ldots, u_H)$ 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 $a_t$ 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 $H$ 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 $K \subseteq \mathcal{V}^*$ is again a sequence of tokens from the common vocabulary $\mathcal{V}$ and $K$ 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.

## 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 fine heuristics to select responses (Stasaski et al., 2022), making large pretrained models (Tack and Piech, 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 \( y \) given a tutoring session \( H = (u_1, \ldots, u_{|H|}) \). 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 \( a = \{a_1, \ldots, a_{|H|}\} \) to be given as an input. Then, given dialog history \( H_{<t} = \{u_1, \ldots, u_t\} \), grounding information \( K \), and \( a_{t+1} \subseteq A_{\text{teacher}} \), the set of dialog acts relevant at timestep \( t + 1 \), the teacher response \( y \) 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.

\[
\begin{align*}
y^* & = \arg\max_{y \in V^*} \{ p(y \mid a_{t+1}, H_{<t}, K) \} \\
& = \arg\max_{y \in V^*} \prod_{i=1}^{|y|} \{ p(y_i \mid y_{<i}, a_{t+1}, H_{<t}, K) \}
\end{align*}
\]

We separate the turns in the dialog by special \( \langle \text{teacher} \rangle \) and \( \langle \text{student} \rangle \) tags and prepend the dialog act as a special token, followed by a special \( \langle \text{knowledge} \rangle \) 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
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.

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

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 e, the model from equation 1 becomes:

\[ p(\mathbf{y} | \mathbf{a}_{t+1}, \mathbf{c}, \mathcal{H}_c, K) \]

### 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 \( a_{t+1} \in A_{\text{teacher}} \) (instead of assuming the ground-truth dialog act) and then uses it to generate a response \( y = u_{t+1} \). 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 \( \tilde{y} := a_{t+1} \circ y \) with act sequence \( a_{t+1} \) of length \( N \) and response \( y \) of length \( T \), the model is

\[ p(\tilde{y} | K, \mathcal{H}_c) = \prod_{i=1}^{m+N} p(\tilde{y}_i | \tilde{y}_{<i}, K, \mathcal{H}_c) \]

In training, we use teacher forcing and prepend \( a_{t+1} \) to \( y \) 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 \( \text{enc}_{\mathcal{H}_c} \) and a response encoder \( \text{enc}_{y} \) encode context \( \mathcal{H}_c \) and possible responses \( y \) into a fixed size vector of same dimension \( n \). In our experiments, the weights \( \theta \) of both encoders are shared.

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

\[ \mathcal{L}(\theta; \mathcal{H}, \tilde{y}, \tilde{y}) = |m + d(\text{enc}_{\mathcal{H}_c} \theta(\mathcal{H}), \text{enc}_y \theta(\tilde{y}))| - d(\text{enc}_{\mathcal{H}_c} \theta(\mathcal{H}), \text{enc}_y \theta(\tilde{y}))|^+, \]

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 \( \mathcal{H}_c \), we choose a response \( y^* \) from the training set \( D \) by maximum inner product search using the decision rule

\[ y^* = \arg\max_{y \in D} \{ \text{enc}_{\mathcal{H}_c}(\mathcal{H}_c)^T \text{enc}_y(y) \}. \]

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), DialogGPT (Zhang et al., 2020), T5 (Raffel et al., 2020) and its multilingual version mT5 (Xue et al., 2021).
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), \(^\dagger\): 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 \(3.25 \times 10^{-5}\), 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 \(y\) 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).
Table 3: F1 score of the dialog act classification based on the generated responses of our models.

<table>
<thead>
<tr>
<th>Method</th>
<th>GT</th>
<th>BART_{base}</th>
<th>BART_{large}</th>
<th>CTRL</th>
<th>Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>DA F1</td>
<td>78.3</td>
<td>81.0</td>
<td>70.1</td>
<td>63.0</td>
<td>43.1</td>
</tr>
</tbody>
</table>

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 DialogGPT 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 occurring 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.

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
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.

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_{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.

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

<table>
<thead>
<tr>
<th>Quality Attribute</th>
<th>sBLEU</th>
<th>BERTScore</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency</td>
<td>0.14</td>
<td>0.12</td>
</tr>
<tr>
<td>Coherence</td>
<td>0.17</td>
<td>0.26</td>
</tr>
<tr>
<td>Correctness</td>
<td>0.06</td>
<td>0.15</td>
</tr>
<tr>
<td>Equitable Tutoring</td>
<td>0.08</td>
<td>0.16</td>
</tr>
</tbody>
</table>

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


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


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.
<table>
<thead>
<tr>
<th>Context</th>
<th>Generated teacher response</th>
</tr>
</thead>
<tbody>
<tr>
<td>T: Please try to fill in the blank in Italian. S: how do you say next? T: Remember that &quot;is next to the&quot; is &quot;e vicino&quot; S: I don’t know how to say purple. Is it viola?</td>
<td>Viola is the word for purple.</td>
</tr>
<tr>
<td>T: Please try to fill in the blank in Italian. S: how do you say next? T: Remember that &quot;is next to the&quot; is &quot;e vicino&quot; S: Purple is viola, right?</td>
<td>That is correct.</td>
</tr>
<tr>
<td>T: Please try to fill in the blank in Italian. S: how do you say next? T: Remember that &quot;is next to the&quot; is &quot;e vicino&quot; S: I think purple could be violett.</td>
<td>Viola is the word for purple.</td>
</tr>
<tr>
<td>T: Please try to fill in the blank in Italian. S: how do you say next? T: Remember that &quot;is next to the&quot; is &quot;e vicino&quot; S: I can’t remember the color. Can you give me a hint?</td>
<td>Purple is viola.</td>
</tr>
<tr>
<td>T: Please try to fill in the blank in Italian. S: how do you say next? T: Remember that &quot;is next to the&quot; is &quot;e vicino&quot; S: I guess it is e vicino all’albero. Is it correct to start the sentence with the preposition?</td>
<td>The preposition for &quot;purple&quot; is &quot;viola&quot; Understanding - Meaning mismatch</td>
</tr>
<tr>
<td>T: &quot;Blue&quot; is &quot;blu&quot;. 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)</td>
<td>Almost, tree is all’albero’. Please try again. Understanding - Incorrect solution assessment</td>
</tr>
<tr>
<td>T: &quot;Blue&quot; is &quot;blu&quot;. 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)</td>
<td>Almost, you need to add blu after &quot;all’albero&quot; to make the word &quot;tree&quot;. Understanding - Incorrect translation</td>
</tr>
<tr>
<td>T: &quot;Blue&quot; is &quot;blu&quot;. 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)</td>
<td>Almost, tree is all’albero not albero.</td>
</tr>
<tr>
<td>T: &quot;Blue&quot; is &quot;blu&quot;. 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)</td>
<td>Almost. Al cane is not the correct word for tree.</td>
</tr>
<tr>
<td>T: &quot;Blue&quot; is &quot;blu&quot;. 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)</td>
<td>The correct answer is e di fronte all’albero blu. Pedagogy - Information reveal</td>
</tr>
</tbody>
</table>

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.