📖 Book2Dial: Generating Teacher-Student Interactions from Textbooks for Cost-Effective Development of Educational Chatbots

Junling Wang[1,2] **Jakub Macina**[1,2] **Nico Daheim**[3]
Sankalan Pal Chowdhury[1] **Mrinmaya Sachan**[1]

[1]Department of Computer Science, ETH Zurich
[2] ETH AI Center
[3] Ubiquitous Knowledge Processing Lab (UKP Lab), Department of Computer Science and Hessian Center for AI (hessian.AI), TU Darmstadt

Abstract

Educational chatbots are a promising tool for assisting student learning. However, developing effective chatbots in education has been challenging, as high-quality data is seldom available in this domain. In this paper, we propose a framework for generating synthetic teacher-student interactions grounded in a set of textbooks. Our approaches capture one aspect of learning interactions where curious students with partial knowledge interactively ask a teacher questions about the material in the textbook. We highlight various quality criteria such dialogues should fulfill and compare several approaches relying on either prompting or fine-tuning large language models. We use synthetic dialogues to train educational chatbots and show the benefits of further fine-tuning in different educational domains. However, human evaluation shows that our best data synthesis method still suffers from hallucinations and tends to reiterate information from previous conversations. Our findings offer insights for future efforts in synthesizing conversational data that strikes a balance between size and quality.

https://github.com/eth-lre/book2dial

1 Introduction

Educational chatbots offer a scalable way to improve learning outcomes among students (Kuhail et al., 2023). However, building educational chatbots has been challenging as high-quality data involving teachers and students is difficult to obtain due to various practical reasons such as privacy concerns (Macina et al., 2023a). In response to this, we study the task of generating synthetic teacher-student interactions from textbooks. We create a novel dataset of textbooks drawn from an open publisher of student textbooks and present a framework (📖 Book2Dial) to generate synthetic teacher-student interactions from these textbooks.

Our teacher-student interactions take the form of conversational question-answering (QA) interactions (Choi et al., 2018; Reddy et al., 2019) where curious students ask teachers questions about the textbook and teachers answer these questions based on the textbook. Our approach primarily extends the direct instruction phase of teaching factual knowledge by facilitating more interactive information exchanges with student (Chi and Wylie, 2014). This differs from techniques focusing on teaching using scaffolding (Macina et al., 2023a) or Socratic questioning (Shridhar et al., 2022) which leads to deeper and wider human learning. However, the task of generating high-quality synthetic dialogue data is difficult (Dai et al., 2022). This is amplified in education, where the generated interactions should cover the teaching material in an informative and coherent way (Agrawal et al., 2012; Crosby, 2000). Therefore, it is important to have quality controls on such data, because students might otherwise receive wrong feedback, which could be detrimental to learning.

Thus, in this work, we also formulate some criteria that measure the quality of educational dialogues. For example, it is crucial that the chatbot does not provide students with incorrect information and stays grounded in the textbook, ensuring factual consistency with the knowledge taught. This is particularly important given that large language models (LLMs) are prone to 'hallucinations' or generating plausible but incorrect or unverified information (Ji et al., 2023). While a simple teacher strategy would be to answer with extracted passages from the textbook, this might hurt the coherence of the dialogue which is present in interactive educational situations (Baker et al., 2021). The teacher's response should both be relevant to the student's question (Ginzburg, 2010), as well as, informative as this ensures that key information from the textbook is covered in the dialogue (Tan et al., 2023). We formalize these requirements into

Findings of the Association for Computational Linguistics ACL 2024, pages 9707–9731
August 11-16, 2024 ©2024 Association for Computational Linguistics

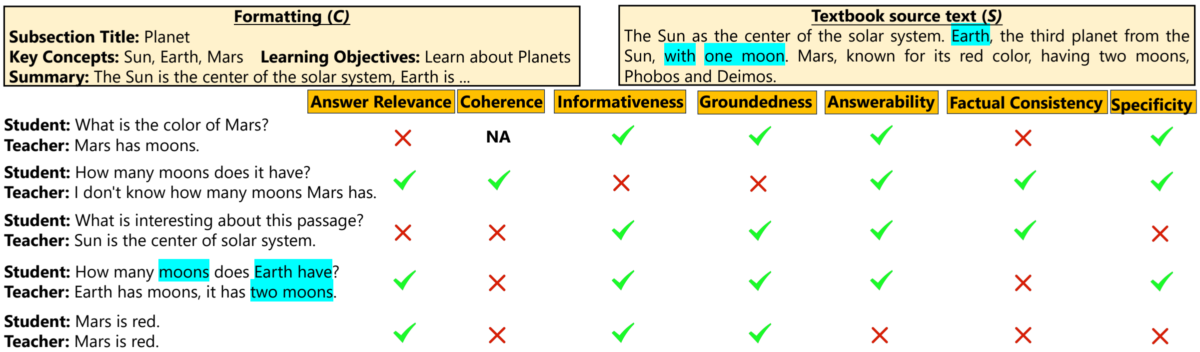

Figure 1: Example of a synthetic teacher-student interaction based on a textbook, along with various criteria for evaluating the quality of the interaction. The criteria include *Answer Relevance* of the answer to the question, *Coherence* of the question-answer interaction with respect to the dialogue history, *Informativeness* of the overall interaction, *Groundedness* to the textbook, *Answerability* of the question from the textbook, *Factual Consistency* of the answer with respect to the question, and *Specificity* of the question. More details in Section 3.2.

criteria shown in Figure 1.

📖 Book2Dial comprises of three approaches taken from prior work: multi-turn QG-QA (Kim et al., 2022), Dialogue Inpainting (Dai et al., 2022) and prompting LLMs to role-play teacher and student. We use the formatting information in the textbook, such as titles or key concepts, to initialize student models with imperfect information. In contrast, the teacher models have perfect information and are expected to generate grounded responses based on the textbook. We fine-tune and prompt various open-source language models to generate teacher-student interactions.

We evaluate 📖 Book2Dial on the proposed quality criteria and use human evaluations to support our findings. Results reveal that data generated by role-playing LLMs scores highest in most criteria, as shown in Section 5.1.1 and 5.1.2, demonstrating reasonable efficacy in creating educational dialogues. The generated dialogues effectively incorporate textbook content but fall short in mimicking the natural scaffolding of educational conversations and exhibit issues like hallucinations and repetition, as discussed in Section 5.3. Despite these limitations, we show that the generated synthetic data may be used to fine-tune educational chatbots with benefits in some educational domains (Section 5.4), especially when the domain of the evaluation dataset matches with the fine-tuning data.

2 Related Work

2.1 Synthetic Data for Conversational QA

Prior work in educational research has focused on generating individual questions (Kurdi et al., 2020) under two common settings: answer-aware and answer-unaware generation. The former approach starts by identifying an answer and then generates a question accordingly, whereas the latter generates a question without pre-determining the answer. These approaches have also been extended to generating multiple questions (Rathod et al., 2022), causal question generation (Stasaski et al., 2021), prediction of question types to ask (Do et al., 2023), or decomposing problems into Socratic subquestions (Shridhar et al., 2022). However, most works do not address conversational settings.

Datasets like QuAC (Choi et al., 2018), CoQA (Reddy et al., 2019) and UltraChat (Ding et al., 2023) focus on conversational question answering in non-educational settings. Previous work has also explored strategies for creating such data with humans or automatically by using models. For example, Qi et al. (2020) withholds the context required for answers from the questioner, leading to information-seeking questions. SimSeek (Kim et al., 2022) synthesizes datasets for conversational question answering from unlabeled documents but does not demonstrate significantly improved performance in downstream tasks. A recent work, Dialogizer (Hwang et al., 2023), proposes a framework for generating context-aware conversational QA dialogues. However, these methods do not take into account the specifics of the educational domain.

Other works outside of conversational question answering also explore data synthesis. For example, Kim et al. (2023) source dialogue data from a commonsense knowledge graph, Chen et al. (2023) prompt LLMs, and Bao et al. (2023) create

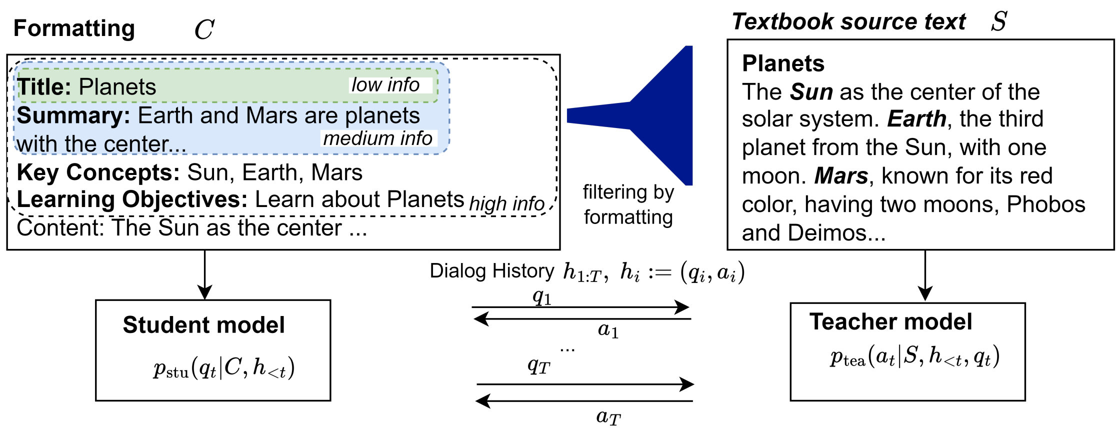

Figure 2: 🕮 `Book2Dial` Framework for Generating Dialogues from Textbooks: Our approach uses two models – a Student model and a Teacher model. The Student model plays the role of a student, formulating questions from a limited context (document formatting). In contrast, the Teacher model assumes the role of a teacher, providing answers and guidance by referencing the (sub-)section in the textbook. This framework can be adapted to various instantiations of the two roles with varying formatting information, such as multi-turn QA-QG models (Kim et al., 2022), Dialogue Inpainting (Dai et al., 2022), and a prompting approach utilizing role-playing LLMs.

grounded dialogue datasets.

2.2 Educational Dialogue Datasets

The development of educational chatbots is highly reliant on high-quality data. Yet such data is hard to obtain. Therefore, previous works such as MathDial (Macina et al., 2023a) collect conversational data by pairing real teachers with an LLM that simulates students. Other datasets are commonly created by roleplaying both teacher and student, such as CIMA (Stasaski et al., 2020) or by transcribing classroom interactions (Suresh et al., 2022; Demszky and Hill, 2023) or logging online conversations (Caines et al., 2020). However, all of these methods are challenging to scale, and using non-experts often leads to data quality issues (Macina et al., 2023a).

Thus, in this work, we explore data synthesis as a scalable way of creating such data. Data augmentation and synthetic data generation gain attention as effective techniques to overcome the challenges associated with manual data annotation. Synthetic data generation is shown to be a promising approach. For instance, Kim et al. (2023) demonstrates the potential of sourcing dialogue data from common sense knowledge. However, ensuring the objectivity of generated data remains a concern. Similarly, Zhang et al. (2018) introduces methods for task-oriented dialogue synthesis. However, its dependency on predefined schemas limits its scalability.

3 Educational Conversation Generation

We first introduce a framework for dialogue synthesis from textbooks in Section 3.1, and then discuss the quality criteria that the generated dialogues should fulfill in Section 3.2.

3.1 🕮 `Book2Dial` Framework

We set out to create meaningful teacher-student interactions from educational textbooks in the form of conversational QA pairs between the teacher and the student. To generate these interactions, we assume that the "teacher" is knowledgeable about the textbook content, and the "student" only knows limited information from the textbook. Thus, we provide the **teacher model** all the textbook information but withhold some information from a **student model**. For this, we can use the structuring and formatting elements found in textbooks, including 1) **Titles:** headings of sections and subsections; 2) **Summary:** summaries of chapters; 3) **Other Metadata:** key concepts, learning objectives, bold terms, and the introductory paragraph of each section; and assume that the student model only has access to this information.

During the conversation, the "student" asks inquisitive questions about the textbook while the "teacher" guides them by answering these questions and including additional information in their response. Formally, a dialogue d comprises of a sequence of T question-answer interactions: $d = \{(q_1, a_1), \ldots, (q_T, a_T)\}$. The formalization of the task is depicted in Figure 2. The student model $p_{\text{stu}}(q_t|C, h_{<t})$ generates a question q_t given the dialogue history $h_{<t} = \{(q_i, a_i)\}_{i=1}^{t-1}$ and the partial context (formatting information) C. The teacher model $p_{\text{tea}}(a_t|S, h_{<t}, q_t)$ generates the answer response a_t given the question, the dialogue history, and the full textbook source S.

3.2 Evaluation of Educational Conversations

To build a high-quality educational conversation, we focus on the student asking questions that are **specific** enough to drive the conversation forward, and also **answerable** given the context. The teacher must then respond with an answer that is **relevant** to the question, **factually consistent** with the context, and **informative** to the student. Finally, the overall conversation should be **coherent** and **grounded** to the entire context, not just parts of it. We use this as our guiding principle and define seven criteria to evaluate the quality of a good educational interaction. Although these metrics may not necessarily be mutually exclusive or extensively address all possible features of high-quality educational dialog, each of them serves as an important aspect to be considered in the education domain. We selected those that could be instantiated by measures and are potentially useful to supplement the traditional factual classroom listening to a lecture with interactive chatbot interactions (Chi and Wylie, 2014; Ruan et al., 2019). We detail these criteria in the rest of this subsection.

3.2.1 Answer Relevance

Answer Relevance measures how directly related the answer is to the question in each QA pair in the dialogue. This criterion is crucial in education, as teachers should have subject-matter knowledge and can adapt responses to the student's various knowledge levels (Lepper and Woolverton, 2002).

We assess the Answer Relevance of individual QA pairs and then combine these assessments to determine the dialogue's overall Answer Relevance. We use **BF1(q_t, a_t)**, **QuestEval** and **Uptake** as metrics for Answer Relevance. The BF1 is an abbreviation for the BERTScore F1 (Zhang* et al., 2020) for semantic alignment between questions and answers using BERT embeddings. QuestEval (Scialom et al., 2021) generates questions from both the question and answer, then generates answers for these questions, comparing them to measure relevance. The Uptake metric (Demszky et al., 2021), specific to the education domain, analyzes teachers' responses to student utterances, focusing on their dependence and relevance. More details are in Table 5.

3.2.2 Coherence of the Dialogue

Coherence measures whether QA pairs in the dialogue form a logical and smooth whole, rather than independent QA pairs. Coherence is an important aspect of good dialogue (Dziri et al., 2019) and it is important in education to sequentially instruct students with the new textbook knowledge in a coherent way (Agrawal et al., 2012).

We adapt two metrics, **BF1(q_t, $a_{<t}$)** and **BF1(q_t, $a_{(t-1)}$)**, to measure coherence, similar to the approach in (Kim et al., 2022). The first metric uses BERTScore F1 to evaluate the current question against each previous answer as references, whereas the second metric compares the current question solely with the immediately preceding answer using BERTScore F1. See Table 5 for more details.

3.2.3 Informativeness

Informativeness evaluates the amount of new information introduced by each student-teacher interaction in the dialogue. This criterion is important in education because the role of providing information is a key aspect of teachers' responsibilities (Crosby, 2000). Informativeness is important to analyze the extent at which the textbook content is progressively introduced to the student.

To assess Informativeness, we use **1 - Overlap(a_t, $a_{<t}$)**, calculating one minus the ratio of token intersection over union in the current and all previous answers for each QA pair. This metric is correlated with human evaluation, as detailed in Appendix E and Table 5.

3.2.4 Groundedness to the Textbook

This criterion assesses the amount of information from the textbook incorporated into the dialogue. Two metrics are used for assessment: **Density**, evaluating the average length of text spans extracted from textbook content S and included in the dialogues; **Coverage**, measuring the proportion of dialogue words originating from the textbook. Both metrics are adopted from (Grusky et al., 2018), and their formulas are shown in Table 5.

3.2.5 Answerability of the Questions

Answerability measures whether the student's question is answerable given the textbook content.

While non-answerable questions could also represent the curiosity of a student (Scialom and Staiano, 2020), the answerability of questions given the context is generally important for a more useful dialog. We use the QA model[1] to judge whether each question is answerable given the textbook content. We refer to this metric as **Answerability**. This approach is akin to the method employed in (Kim et al., 2022). More details are in Table 5.

3.2.6 Factual Consistency of the Answer

Factual Consistency measures whether the answer correctly responds to the student's question. This criterion is crucial in education because it is important for students to learn accurate information (Metzger et al., 2003). Existing metrics like Q^2 (Honovich et al., 2021) use a QA model to assess answer correctness, while RQUGE (Mohammadshahi et al., 2023) uses a QA model to evaluate the quality of the candidate question. In our scenario, we need to measure whether the answer contains correct information and accurately answers the question. Therefore, we build on the idea of Q^2 and introduce a new metric referred as **QFactScore**:

$$\alpha \cdot \mathrm{sim}(\mathrm{QA}(q_t, S), a_t) + \beta \cdot \mathrm{sim}(q_t, a_t) \quad (1)$$

It calculates the cosine similarity of embeddings between the predicted and original answers for each QA pair and also evaluates the similarity between the question and the original answer. This metric has been validated for its alignment with human evaluation, as detailed in Appendix E. The final score is the weighted sum of two similarity scores. More details are in Table 5 and Appendix C.4.

3.2.7 Specificity of the Question

Specificity assesses whether the question is specific, rather than general. An example of a generic question is 'What is interesting about this passage?'. The Specificity criterion is crucial in education, as teachers should ask specific questions (Yang, 2017). We assess specificity through human evaluation, as there is no existing metric that captures specificity.

4 From Textbooks to Dialogues

In this section, we describe different methods used for generating dialogues from educational textbooks in `Book2Dial`, namely:

1. **Multi-turn QG-QA models.** In this setting, we use fine-tuned QG and QA models interacting with each other.
2. **Dialogue Inpainting.** Dai et al. (2021) uses a span extraction model over the textbook as a teacher model, where the response is copied from the textbook and the question is generated by a QG model acting as the student.
3. **Persona-based Generation.** This approach uses LLMs like GPT-3.5, and leverages prompting to interactively simulate the student and the teacher and generate dialogues.

We describe these methods next. More implementation details are in Appendix C.

4.1 Multi-turn QG-QA models

This scenario utilizes separate QG and QA models to interact in a multi-turn scenario. As a representation of this approach from related work, we consider the SimSeek-asym model (Kim et al., 2022). The approach consists of two components:

1. A **Question Generation** (QG) model for generating conversational questions relying solely on prior information (i.e., formatting information relevant to the topic). This student model generates a question based on the dialogue history and filtered Information C: $p(q_t|C, h_{<t})$.
2. A **Conversational Answer Finder** (CAF) to comprehend the generated question and provide an acceptable teacher answer to the question from the evidence passage: $p(a_t|S, h_{<t}, q_t)$.

4.2 Dialogue Inpainting

Dialogue Inpainting (Dai et al., 2022) is an approach for dialogue generation characterized by its information-symmetric setting. In this framework, both the student and teacher model are provided with the complete textbook text S. The teacher model iterates over each sentence in S and copies it as an answer. The student model is a QG model. We use data from the OR-QuAC (Qu et al., 2020), QReCC (Anantha et al., 2021), and Taskmaster-2 (Byrne et al., 2020) datasets to train the student model. For the student model, a dialogue reconstruction task is employed. At training time rather than distinguishing questions and answers, the dialogue reconstruction task treats a conversation as a sequence of utterances $\{u_i\}_{i=1}^{2T}$, To train it, a randomly chosen utterance u_i is masked to create a

[1]https://huggingface.co/distilbert-base-cased-distilled-squad

partial dialogue $d_{m(i)} = u_1, \cdots, u_{i-1}, < mask >, u_{i+1}, \cdots, u_{2T}$. The model then predicts u_i and is trained by minimizing the loss:

$$\mathcal{L}(\theta) = -\sum_{d \in D} \mathbb{E}_{u_i \sim d}[\log p_\theta(u_i \mid d_{m(i)})] \quad (2)$$

During inference, the model uses each sentence in the textbook as a teacher's utterance and only predicts student utterances accordingly, $\{u_{2k-1}\}_{k=1}^{T}$ corresponding to $\{q_i\}_{i=1}^{T}$ in our notation. We basing our model (eq 2) on FLAN-T5-XL (Chung et al., 2024). More details are in Appendix C.2.

4.3 Persona-based Generation

Inspired by (Markel et al., 2023)'s idea of using LLMs to simulate student personas, we propose a method to simulate student and teacher personas by prompting LLMs for dialogue generation. We use one instance of the GPT-3.5 model to play the student and another to play the teacher.[2] The teacher model is provided with all the information from the textbook, including content and all the formatting information. The information provided to the student model is varied. We consider four variants for generating dialogue in each subsection listed with an increasing amount of information provided to the student model: 1) *Persona (Low Info)* provides the student model with only the Title information, 2) *Persona (Medium Info)* provides both the Title and Summary information, 3) *Persona (High Info)* offers all formatting information, and 4) *Persona (Single Instance)* generates the entire dialogue using a single LLM instance, equipping one model with formatting and textbook content information. More details are in Table 7.

Considering that GPT-3.5 is not open-source, we use prompting to steer the models in dialogue generation. The prompt for Persona (High Info) and Persona (Single Instance) is detailed in Appendix C.3.

5 Results and Analyses

In this section, we aim to address the following research questions:

1. How does the choice of generation framework influence the quality of the generated data?
2. What is the optimal amount of information that should be incorporated into the student model to produce natural dialogues?
3. How does fine-tuning on the synthesized data in various domains improve the performance of models that are fine-tuned and evaluated on existing datasets?

To address these questions, we generate dialogues from textbooks across various domains and analyze the generated dataset.

Textbook data: We collect 35 textbooks available on OpenStax[3], spanning domains of math, business, science, and social science. From these, we select four textbooks to create our dialogue datasets. Table 6 provides statistics of the four textbooks. The first and second research questions are addressed in Sections 5.1 and 5.2, respectively, while the third question is answered in Section 5.4.

5.1 Automatic Evaluation

In this section, we discuss statistics and metrics for the generated datasets. We present aggregated statistics of the datasets across four textbook domains in Tables 1 and 2, also noting comparisons with existing datasets: MathDial (Macina et al., 2023a) – math tutoring conversations between human teachers and student LLMs, fact-based human-human conversations of QuAC (Choi et al., 2018), and NCTE transcripts of math classrooms (Demszky and Hill, 2023). Domain-specific results in the dataset are detailed in Tables 10 and 11. To adjust for varying dialogue lengths, we limit the number of turns to $T = 12$ for each model, as in (Kim et al., 2022).

5.1.1 Statistical Analysis

In dialogue, different types of questions emphasize various aspects. We hypothesize that "what" and "which" questions focus on factual information. In contrast, other question types, such as "why" and "how," tend to reflect more complex inquiries, which are also important in educational contexts. In Table 1, we present the percentages of student questions including words *what*, *which*, *why*, and *how*[4]. Furthermore, the average token count for questions and answers across each dataset is also shown. The key findings are as follows:

[2]We used the GPT-3.5-turbo API between 25th September and 4th October, 2023.

[3]https://openstax.org/

[4]This ratio excludes 'how much' and 'how many' questions because they pertain to factual information.

	Question Type (%)			Num. Tokens in:	
	what / which	why	how	Questions	Answers
SimSeek	55.00	2.00	17.50	10.90	14.33
Dialogue Inpainting	63.50	2.75	16.00	6.85	19.63
Persona (Single Inst.)	28.25	**5.50**	23.25	14.97	34.58
Persona (Low Info)	70.25	0.75	**27.50**	17.56	84.75
Persona (Med. Info)	69.00	0.25	27.00	17.69	85.19
Persona (High Info)	**73.50**	0.75	24.00	19.01	84.70
MathDial	21.00	5.00	10.00	17.11	32.91
QuAC	36.00	3.00	8.00	6.52	12.62
NCTE	40.00	4.00	10.00	33.85	4.41

Table 1: Key statistics of the synthesized educational dialogue dataset: Persona-generated datasets contain more 'what/which' and 'how' questions and the dialogues tend to be more verbose.

Less factual questions in the Persona (Single Instance) dataset The Persona (Single Instance) model generates the fewest "what" or "which" questions compared to other synthesized datasets, suggesting more diverse questioning. It also has a similar question type distribution to MathDial.

More "how" questions in Persona datasets The four Persona datasets contain the highest ratio of 'how' questions, which suggests a higher ratio of questions asking for explanations.

High token counts in Persona datasets Datasets from Persona models feature the high average token counts in questions and answers, suggesting these dialogues are more verbose and informative. NCTE features high token counts in questions, typical in classroom transcripts with lengthy teacher inquiries and brief student responses.

5.1.2 Data Quality Metrics

We report the various data quality metrics in Table 2. Our key findings are as follows:

Persona datasets are solid in most of the criteria Persona models generated datasets outperform others in most metrics, indicating their good ability to create dialogues from textbooks. Note the MathDial, QuAC, and NCTE datasets each focus on different domains with different data collection designs and grounding texts, therefore they are not directly comparable in these metrics and should be interpreted with their specific context in mind. For example, a tutoring dataset like MathDial is expected to have lower answerability as tutors tend to ask challenging questions while the specific data collection of QuAC to include supporting arguments increases the density.

High Informativeness and Groundedness of Dialogue Inpainting dataset: Dialogue Inpainting models achieve the highest score across all models in Informativeness and Groundedness. This is expected as this model uses exact sentences from the textbooks as teachers' answers.

Students with more information access perform better in automatic metrics. Datasets from Persona (High Info) and Persona (Medium Info) typically outperform or match with datasets from Persona (Low Info). This suggests that more information to the student may enhance the key criteria. However, the impact differences among formatting levels are not markedly significant, indicating a need for further research on this question.

5.2 Human Evaluation

To compensate for the limitations of automatic metrics, we conduct human evaluations of SimSeek, Dialog Inpainting, and Persona (High Info) dialogues based on seven criteria: Answer Relevance (AnsRel), Informativeness (Info), Groundedness (Gro), Coherence (Coh), Factual Consistency (Fact), Answerability (Ans), and Specificity (Spe). Questions for judging each criterion are in Table 12. We recruited 4 annotators to evaluate 12 dialogues each, yielding a Fleiss' kappa of 0.74, indicating reasonable agreement. Evaluation details are in Appendix F, and results in Table 3.

Persona (High Info) excels among the three models, leading in Answer Relevance, Coherence, Factual Consistency, Answerability, and Specificity, rendering it the most suitable choice based on our dialogue generation objectives. This result aligns with the results of automatic metrics presented in Table 2. However, the dialogues generated by the Persona-based method exhibit only an average score in Informativeness, with a score of 0.74 indicating that approximately 26% of QA pairs do not contribute new information to the conversation. The Persona-based model, while leading in Factual Consistency among the three models, scores only 0.79, which indicates that approximately 21% of the QA pairs lack Factual Consistency. For educational dialogues, it's imperative to aim for high Factual Consistency to ensure the reliability of the knowledge imparted. The primary reason for this issue is the hallucination in LLMs, where LLMs respond to questions using fabricated or false information not grounded in the textbook. This poses a significant challenge and calls for further research

	Answer Relevance			Informativeness	Groundedness		Coherence		Answerability	Factual Consistency
	BF1 (q_t, a_t)	QuestEval	Uptake	1 - Overlap $(a_t, a_{<t})$	Density	Coverage	BF1 $(q_t, a_{<t})$	BF1 (q_t, a_{t-1})	Answerable	QFactScore
SimSeek	0.53	0.25	0.78	0.71	11.66	0.82	0.51	0.55	0.84	0.32
Dialogue Inpainting	0.52	0.28	0.84	**0.91**	**22.62**	**0.90**	0.45	0.46	0.75	0.24
Persona (Sing. Inst.)	0.58	0.35	0.98	0.86	3.94	0.75	0.49	0.52	0.92	0.54
Persona (Low Info)	0.61	**0.44**	**0.99**	0.59	2.39	0.70	0.52	**0.59**	0.98	0.75
Persona (Med. Info)	0.61	**0.44**	**0.99**	0.59	2.43	0.71	0.52	**0.59**	**0.99**	**0.76**
Persona (High Info)	**0.62**	**0.44**	**0.99**	0.60	2.50	0.71	**0.53**	**0.59**	**0.99**	0.75
MathDial	0.46	0.30	0.83	0.64	1.30	0.46	0.42	0.47	0.51	0.39
QuAC	0.43	0.24	0.76	0.72	13.78	0.81	0.42	0.43	0.73	0.38
NCTE	0.34	0.21	0.76	0.89	NA	NA	0.38	0.37	NA	NA

Table 2: Automatic evaluation metrics for the synthesized dialogue data and existing datasets. Higher values indicate better performance. Persona-generated dialogues score highest in Answer Relevance, Coherence, Answerability, and Factual Consistency, while Dialogue Inpainting generated dialogues score highly in Informativeness and Groundedness.

	AnsRel	Info	Gro	Coh	Fact	Ans	Spe
SimSeek	0.32	0.56	**1.00**	0.58	0.25	0.66	0.89
Dial. Inpaint.	0.58	**0.97**	**1.00**	0.66	0.73	0.83	0.61
Persona (High Info)	**0.97**	0.74	0.99	**0.85**	**0.79**	**0.96**	**0.99**

Table 3: Human Evaluation Result: Persona (High Info) generated dialogues score highest in Answer Relevance, Coherence, Factual Consistency, Answerability, and Specificity, while Dialogue Inpainting generated dialogues excel in Informativeness and Groundedness.

into ways to better ground LLMs to text documents in the future.

5.3 Qualitative Human Analysis

We inspect dialogues generated by each model and find several limitations compared to natural educational conversations. We report qualitative descriptions of these findings:

Insufficient follow-up ability of Persona models Even though Persona models outperformed other methods in automatic and human evaluations, the model has several issues. Dialogues generated by Persona models are unlike natural conversations and resemble a series of artificial QA pairs about textbooks. Moreover, the dialogue does not have enough follow-up questions by students and rather broadly touches on the textbook content rather than going into depth about certain topics. We present an example of a dialogue demonstrating the insufficient follow-up ability of the Persona-based model in table 17.

Repeating answers in SimSeek and Persona In the SimSeek and Persona datasets, we find that teacher answers often reiterate information from previous interactions. SimSeek often generates questions related to the previous teacher's answer, while Persona often provides summaries of textbook content in each answer.

Insufficient Specificity of Dialogue Inpainting In alignment with the results of human evaluation, we find that the Dialogue Inpainting model tends to generate "general" questions, such as "What is interesting about this passage?" These types of questions, which are not specific to the textbook content, are less desirable in educational dialogue.

5.4 Fine-tuning for Educational Chatbots

In this section, we show the effectiveness of these imperfect but automatically created synthetic data based on any textbook domain for the downstream task of fine-tuning educational chatbots. We fine-tune simple chatbot models with our synthesized data and assess their performance on educational conversation task of next utterance generation.

Specifically, we use text generation models based on language models to generate teacher responses a_t given the dialogue history $h_{<t}$, textbook grounding information S, and the question q_t. We compare two scenarios: (1) a model fine-tuned on synthetic textbook data, then further fine-tuned and tested on various educational or information-seeking dialogue datasets; and (2) a model fine-tuned and tested solely on these dialogue datasets without fine-tuning on synthetic textbook data. We use FLAN-T5-LARGE (Chung et al., 2024) as our base language model. For our test sets, we use the MCTest and CNN splits of the CoQA dataset (Reddy et al., 2019), as well as the NCTE dataset (Demszky and Hill, 2023). The MCTest split contains dialogues about children's stories; the CNN split contains conversations about the news; the NCTE dataset contains transcripts of elementary math classrooms.

We fine-tune the base model on four textbook-based synthetic datasets, each from a different subject: math, business, science, and social science. The datasets and training details are shown in Appendix G. The results are shown in Table 4. We report the BLEU score for the scenario where we fine-tune the base model on our textbook-generated dialogue dataset, with the difference between this fine-tuned version and the version without fine-tuning shown in brackets. We use two baselines: Zero-shot (0-Shot), which directly uses the FLAN-T5-LARGE model without any fine-tuning, and UltraChat, which is fine-tuned on the synthetic but non-educational UltraChat dataset (Ding et al., 2023) and tested on different target datasets.

We find that the model that is first fine-tuned on the social science textbook data achieves the highest score when tested on the MCTest and CNN splits of the CoQA dataset, with improvements of 4.16 and 1.99. Meanwhile, the model fine-tuned on the business textbook data achieves the highest score when tested on the NCTE dataset. The model fine-tuned on the math textbook data also shows improvements. As the social science textbook dataset contains the fewest math expressions, it improves the most in non-math domains but performs the worst in the math domain. We conclude that synthetic datasets created from textbooks, as well as Ultrachat synthetic data, may be effective for fine-tuning chatbots if they align with the target domain.

Upon a more qualitative human examination of the generated results, we find that the fine-tuned models on synthetic textbook data have a better understanding of the input context and generate more correct answers than the corresponding non-fine-tuned models. Some example generations are shown in Appendix G.1.

6 Conclusion

This work introduces a new task of generating conversational question-answering dialogues from textbooks to help fine-tune educational chatbots in various domains where high-quality dialogue data are scarce. We detail and compare various approaches and settings to simulate student-teacher interactions and create such data. We evaluate the generated dialogues, focusing on some measures of their quality, such as Answer Relevance, Informativeness, Coherence, and Factual Consistency. Our results indicate that the approach with LLMs role-playing as teachers and students for data synthesis performs well in these metrics. However, upon closer human inspection, we also observe several issues with the synthesized data, such as the problem of hallucinations and repeating information. Despite these issues, we show that the generated synthetic dialogues using 📖 Book2Dial can be used to fine-tune educational chatbots and achieve performance improvements in various educational settings.

	CoQA (MCTest)	CoQA (CNN)	NCTE
Math	26.10 (+3.96)	13.95 (+0.82)	8.79 (+0.39)
Business	18.91 (-3.23)	13.29 (+0.16)	**8.99 (+0.59)**
Science	22.36 (+0.22)	14.96 (+1.83)	8.73 (+0.33)
Social	**26.30 (+4.16)**	**15.11 (+1.98)**	8.37 (-0.03)
All	23.05 (+0.91)	14.31 (+1.18)	8.41 (+0.01)
UltraChat	25.47 (+3.33)	14.89 (+1.76)	8.74 (+0.34)
0-Shot	3.96 (-18.18)	3.67 (-9.46)	1.08 (-7.32)

Table 4: Downstream Task Results. We use dialogues generated from one textbook from each domain for fine-tuning and evaluate on downstream benchmarks. Each cell displays the BLEU score and the (difference from the baseline), where the baseline is derived from the same model without fine-tuning using 📖 Book2Dial synthetic data. Two additional baselines are: Zero-shot (0-Shot), which directly uses the FLAN-T5-LARGE model without any training, and UltraChat, which is fine-tuned on UltraChat and tested on different target datasets.

7 Limitations

Focus on a specific teaching scenario and limitations in educational contexts In this work, we focus on a specific educational scenario where a curious student asks questions to a knowledgeable teacher. It is shown that the quality of the student's questions (with deep reasoning ones) is correlated with their learning (Graesser and Person, 1994; Person et al., 1994). We do not model any of these aspects in our approach. Furthermore, recent approaches of teachers asking Socratic questions or providing indirect scaffolds and hints instead of providing students directly with answers are also shown to lead to better learning outcomes (Freeman et al., 2014). In our formulation, teachers directly provide students with answers. Our approach focuses on facilitating informational exchanges and is more suitable for helping students access the entire content of the textbook in a new interactive way and through their interests. This serves as a starting point for developing more sophisticated educational chatbots in the future. Future work could focus on other interaction scenarios and combine our approach with Socratic questioning (Shridhar et al., 2022) and scaffolding (Macina et al., 2023b; Sonkar et al., 2023) to achieve significantly improved applicability to educational use cases.

Achieving the highest scores in all metrics is not the overall goal leading to the most effective human learning. Considering Informativeness, while a dialogue rich in information suggests a potential for a greater extent of learning by a student, there exists a trade-off, as excessive information can increase the student's cognitive load and become overwhelming (Mayer and Moreno, 2003). Therefore, finding the optimal amount of information that the dialogue should contain needs careful consideration in future work. Similarly, for other metrics, educational practitioners could ideally set the target metrics and their combination for achieving better fine-tuning performance steered towards educational use cases.

Aspects of evaluation framework: Although we try to include various aspects of the evaluation in this work, it is not feasible to focus on all important educational aspects. We specifically focus on one setting, where students ask curious questions and the teacher provides answers. Therefore, comparing our datasets with the MathDial, QuAC, and NCTE datasets does not fully explain our datasets' quality, as MathDial, QuAC, and NCTE datasets are focused on different interaction situations. In particular, none of the MathDial, QuAC, or NCTE datasets include textbook content; MathDial contains only short math problems and focuses on scaffolding, while QuAC is oriented towards fact-based queries rather than student-teacher interactions; NCTE consists of classroom transcripts in which there are more than just two interlocutors.

8 Acknowledgements

This project was made possible by ETH AI Center Doctoral Fellowships to Jakub Macina with partial support from the Asuera Stiftung and the ETH Zurich Foundation. Nico Daheim acknowledges the 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 Xiaoyu Zhang for the insightful discussion about the design of the human evaluation experiments and Kumar Shridhar for help in collecting the Openstax textbooks. Additionally, the authors wish to thank reviewers, members of the LRE group at ETH Zurich, and the participants in human evaluation experiments.

References

Rakesh Agrawal, Sunandan Chakraborty, Sreenivas Gollapudi, Anitha Kannan, and Krishnaram Kenthapadi. 2012. Quality of textbooks: An empirical study. In *Proceedings of the 2nd ACM symposium on computing for development*, pages 1–1.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534, Online. Association for Computational Linguistics.

Michael J Baker, Baruch B Schwarz, and Sten R Ludvigsen. 2021. Educational dialogues and computer supported collaborative learning: critical analysis and research perspectives. *International Journal of Computer-Supported Collaborative Learning*, pages 1–22.

Jianzhu Bao, Rui Wang, Yasheng Wang, Aixin Sun, Yitong Li, Fei Mi, and Ruifeng Xu. 2023. A synthetic data generation framework for grounded dialogues. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10866–10882, Toronto, Canada. Association for Computational Linguistics.

Bill Byrne, Karthik Krishnamoorthi, Saravanan Ganesh, Amit Dubey, Andy Cedilnik, and Kyu-Young Kim. 2020. Taskmaster-2. `https://github.com/google-research-datasets/Taskmaster/tree/master/TM-2-2020`. Second dataset in series of three.

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.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. PLACES: Prompting language models for social conversation synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Joy Crosby, RM Harden. 2000. Amee guide no 20: The good teacher is more than a lecturer-the twelve roles of the teacher. *Medical teacher*, 22(4):334–347.

Nico Daheim, Nouha Dziri, Mrinmaya Sachan, Iryna Gurevych, and Edoardo M Ponti. 2023. Elastic weight removal for faithful and abstractive dialogue generation. *arXiv preprint arXiv:2303.17574*.

Shuyang Dai, Guoyin Wang, Sunghyun Park, and Sungjin Lee. 2021. Dialogue response generation via contrastive latent representation learning. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 189–197, Online. Association for Computational Linguistics.

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. In *International Conference on Machine Learning*, pages 4558–4586. PMLR.

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

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

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore. Association for Computational Linguistics.

Xuan Long Do, Bowei Zou, Shafiq Joty, Tran Tai, Liangming Pan, Nancy Chen, and Ai Ti Aw. 2023. Modeling what-to-ask and how-to-ask for answer-unaware conversational question generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10785–10803, Toronto, Canada. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 146–148, Florence, Italy. 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.

Jonathan Ginzburg. 2010. Relevance for dialogue. In *SemDial: Workshop on the Semantics and Pragmatics of Dialogue (PozDial)*, pages 121–129.

Arthur C Graesser and Natalie K Person. 1994. Question asking during tutoring. *American educational research journal*, 31(1):104–137.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. 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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yerin Hwang, Yongil Kim, Hyunkyung Bae, Hwanhee Lee, Jeesoo Bang, and Kyomin Jung. 2023. Dialogizer: Context-aware conversational-QA dataset generation from textual sources. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8806–8828, Singapore. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Gangwoo Kim, Sungdong Kim, Kang Min Yoo, and Jaewoo Kang. 2022. Generating information-seeking conversations from unlabeled documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2362–2378.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. SODA: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949, Singapore. Association for Computational Linguistics.

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

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204.

Mark R. Lepper and Maria Woolverton. 2002. Chapter 7 - the wisdom of practice: Lessons learned from the study of highly effective tutors. In Joshua Aronson, editor, *Improving Academic Achievement*, Educational Psychology, pages 135–158. Academic Press, San Diego.

I Loshchilov and F Hutter. 2019. " decoupled weight decay regularization", 7th international conference on learning representations, iclr. *New Orleans, LA, USA, May*, (6-9):2019.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023a. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore. Association for Computational Linguistics.

Jakub Macina, Nico Daheim, Lingzhi Wang, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023b. Opportunities and challenges in neural dialog tutoring. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2357–2372, Dubrovnik, Croatia. Association for Computational Linguistics.

Julia M Markel, Steven G Opferman, James A Landay, and Chris Piech. 2023. Gpteach: Interactive ta training with gpt-based students. In *Proceedings of the tenth acm conference on learning@ scale*, pages 226–236.

Richard E Mayer and Roxana Moreno. 2003. Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist*, 38(1):43–52.

Miriam J Metzger, Andrew J Flanagin, and Lara Zwarun. 2003. College student web use, perceptions of information credibility, and verification behavior. *Computers & Education*, 41(3):271–290.

Alireza Mohammadshahi, Thomas Scialom, Majid Yazdani, Pouya Yanki, Angela Fan, James Henderson, and Marzieh Saeidi. 2023. RQUGE: Reference-free metric for evaluating question generation by answering the question. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6845–6867, Toronto, Canada. Association for Computational Linguistics.

Natalie K Person, Arthur C Graesser, Joseph P Magliano, and Roger J Kreuz. 1994. Inferring what the student knows in one-to-one tutoring: The role of student questions and answers. *Learning and individual differences*, 6(2):205–229.

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, Brussels, Belgium. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, and Christopher D. Manning. 2020. Stay hungry, stay focused: Generating informative and specific questions in information-seeking conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 25–40, Online. Association for Computational Linguistics.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 539–548.

Manav Rathod, Tony Tu, and Katherine Stasaski. 2022. Educational multi-question generation for reading comprehension. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 216–223.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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*, pages 1–13.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Scialom and Jacopo Staiano. 2020. Ask to learn: A study on curiosity-driven question generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2224–2235, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kumar Shridhar, Jakub Macina, Mennatallah El-Assady, Tanmay Sinha, Manu Kapur, and Mrinmaya Sachan. 2022. Automatic generation of socratic subquestions for teaching math word problems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4136–4149, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shashank Sonkar, Naiming Liu, Debshila Mallick, and Richard Baraniuk. 2023. CLASS: A design framework for building intelligent tutoring systems based on learning science principles. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1941–1961, Singapore. Association for Computational Linguistics.

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.

Katherine Stasaski, Manav Rathod, Tony Tu, Yunfang Xiao, and Marti A Hearst. 2021. Automatically generating cause-and-effect questions from passages. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 158–170.

Abhijit Suresh, Jennifer Jacobs, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022. Finetuning 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.

Wei Tan, Jionghao Lin, David Lang, Guanliang Chen, Dragan Gašević, Lan Du, and Wray Buntine. 2023. Does informativeness matter? active learning for educational dialogue act classification. In *International Conference on Artificial Intelligence in Education*, pages 176–188. Springer.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Hao Yang. 2017. A research on the effective questioning strategies in class. *Science Journal of education*, 5(4):158–163.

Lichao Zhang, Abel Gonzalez-Garcia, Joost Van De Weijer, Martin Danelljan, and Fahad Shahbaz Khan. 2018. Synthetic data generation for end-to-end thermal infrared tracking. *IEEE Transactions on Image Processing*, 28(4):1837–1850.

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

A Metrics Formulas

The metrics mentioned in Section 3.2 are detailed and explained in Table 5, including formulas and explanations.

B Textbook Statistics

The four textbooks we use to generate dialogue for experiments are collected from the OpenStax website. The math textbook is titled 'Introductory Statistics,' the business textbook 'Business Ethics,' the science textbook 'Physics,' and the social science textbook 'Psychology 2e.' The statistics of the four textbooks are shown in Table 6.

C Implementation Details

C.1 Information-seeking scenario

In the SimSeek-ASYM setup, the CQG model ingests the title and summary information, each separated by special tokens. We use T5-Large as the student's model and Longformer-Large as the teacher's model.

The SimSeek-ASYM code[5] can be executed with minor modifications. We use the same CQG and CAF models as in (Kim et al., 2022), which utilize T5 as the student's model and Longformer as the teacher's model.

C.2 Dialogue Inpainting

We adopt a training regimen that integrates data from the OR-QuAC (Qu et al., 2020), QReCC (Anantha et al., 2021), and the movie and restaurant datasets from Taskmaster-2 (Byrne et al., 2020), employing the technique as described in (Dai et al., 2022). We randomly select 80% of the data as the training set, while the remaining 20% as the test set. We implement Dialogue Inpainting using the code framework of (Daheim et al., 2023), basing our model (eq 2) on FLAN-T5-XL (Chung et al., 2024), and train it with LoRA (Hu et al., 2022) to reduce computational load. We set an initial learning rate of 6.25e-5 and employ linear learning rate decay without warmup. For model optimization, we utilize checkpoints from the transformers library (Wolf et al., 2020). The negative log-likelihood of the ground-truth response is minimized using the AdamW optimizer, as detailed in (Loshchilov and Hutter, 2019). We assess model performance using the sacrebleu implementation of the BLEU metric, following (Post, 2018). We use one V100 GPU to train the model. The FLAN-T5-XL model has 3 billion parameters and takes 12 hours to train.

The model, while fundamentally designed to predict single utterances, is used autoregressively. It begins with the input s_{prompt}, $< mask >$, s_1 and sequentially generates questions using top-p sampling. This autoregressive process continues until the dialogue is wholly formed.

[5]https://github.com/naver-ai/simseek

C.3 Persona-based Generation

Prompt for Persona (High Info) The design of our prompts is chiefly driven by the requisites of context-awareness, speaker identification, and specificity. We incorporate guidelines and annotations to ensure GPT yields concise responses and minimizes redundant information. To distinguish between speakers, we prefix dialogues with labels: "Teacher:" or "Student:". The prompt is shown below.

Prompt for simulating student

```
Task: You are a student preparing to
ask questions about a textbook subsection
to a teacher. Your goal is to uncover
the key information from this subsection.
Based on the teacher's responses, you'll
further inquire to get a comprehensive
understanding. Make sure to ask specific
questions about the subsection's content
and avoid repeating queries from prior
discussions.
Information Provided:

 1. Section Title: ...

 2. Subsection Title: ...

 3. Section Summary: ...

 4. Bold Terms in Section: ...

 5. Learning Objectives: ...

 6. Concepts in Section: ...

 7. Section Introduction: ...

Previous Conversation:

    Student:...

    Teacher:...

*Note:* Frame your questions considering
the information above and ensure they're
relevant to the content.  Do not ask
```

Criterion	Metric	Definition	Explanation
Answer Relevance	BF1 (q_t, a_t)	BERTScoreF1(q_t, a_t)	For each QA pair, we compute the BERTScore F1(Zhang* et al., 2020), treating the question as the predicted sentence and the answer as the reference sentence. It evaluates the semantic correspondence between the question and answer using BERT's contextual embeddings.
	QuestEval	QuestEval(q_t, a_t)	For each QA pair, we compute the QuestEval score (Scialom et al., 2021), treating the question as the predicted sentence and the answer as the reference sentence. QuestEval generates questions from both the original question and the answer, then generates answers for these questions, comparing their consistency and completeness to evaluate Answer Relevance.
	Uptake	Uptake(q_t, a_t)	For each QA pair, we compute the Uptake score (Demszky et al., 2021) between student and teacher utterances. Uptake is computed as pointwise Jensen-Shannon Divergence (pJSD), estimated through next utterance classification, to analyze the teacher's responses to student utterances in terms of their dependence and relevance.
Coherence	BF1 $(q_t, a_{<t})$	BERTScoreF1$(q_t, a_{<t})$	It computes the BERTScore F1 for each dialogue question, treating it as the predicted sentence against all preceding answers as references. Aggregated scores reflect the dialogue's coherence.
	BF1 $(q_t, a_{(t-1)})$	BERTScoreF1$(q_t, a_{(t-1)})$	It computes the BERTScore F1 for each dialogue question against the immediately preceding answer as the reference. Aggregated scores provide a measure of overall coherence.
Informative- -ness	1-Overlap $(a_t, a_{<t})$	$1 - \frac{\lvert\mathbf{a}_t \cap \mathbf{a}_{<t}\rvert}{\lvert\mathbf{a}_t \cup \mathbf{a}_{<t}\rvert}$	For each answer in a dialogue, the proportion of its intersection with previous answers to their union is computed using word-level tokens. This value is then subtracted from 1.
Content Match	Density	$\frac{1}{\lvert h_{1:T}\rvert} \sum_{f \in \mathcal{F}(S, h_{1:T})} \lvert f\rvert^2$,$\mathcal{F}(S, h_{1:T})$:the set of extractive phrases in dialogue $h_{1:T}$ extracted from textbook content S.	Density refer to Extractive Fragment Density (Grusky et al., 2018), as the average length of text spans that are directly extracted from textbook content S and included in the dialogues.
	Coverage	$\frac{1}{\lvert h_{1:T}\rvert} \sum_{f \in \mathcal{F}(S, h_{1:T})} \lvert f\rvert$	Coverage refer to Extractive Fragment Coverage (Grusky et al., 2018), as the percentage of words in a dialogue that originated from the textbook content.
Answerability	Answerable	Valid(QA(q_t,S))	We use the "distilbert-base-cased-distilled-squad" QA model to determine if a question is answerable from the textbook content. If it generates an empty string or an invalid answer such as "CANNOTANSWER", the question is deemed unanswerable. We report the ratio of answerable questions as 1 minus the ratio of unanswerable questions.
Factual Consistency	QFactScore	$\alpha sim(QA(q_t, S), a_t) + \beta sim(q_t, a_t)$	For each QA pair, it computes the cosine similarity between the embeddings of the QA model's predicted answer and the original answer. Then, it assesses the similarity between the embeddings of the question and answer. The final score is the weighted sum of two similarity scores.
Specificity	NA	NA	We lack automatic metrics for evaluating this criterion.

Table 5: Formulas and Explanations of Each Metric: except 1-Overlap and QFactScore, all metrics are adopted from previous research.

Domain	Name	Chapters	Paragraphs	Pages	Words
Math	Introductory Statistics	13	1,412	65	35,182
Business	Business Ethics	11	795	42	85,626
Science	Physics	23	1,918	89	106,712
Social science	Psychology 2e	16	1,710	88	191,273

Table 6: Summary of the textbook statistics: The Social Science textbook is the longest, while the Math textbook is the shortest.

`questions about information you already have. Only ask one question at a time. Expected Output: Please phrase your question as a string.`

Prompt for simulating teacher `Task: You are a teacher preparing to answer a student's question about a subsection of a textbook. The student's question is: {question}. Provide a concise, specific response, ensuring it's not a summary and distinct from any previous answers you've given.`

```
Information Provided:
 1. Section Title: ...
 2. Subsection Title: ...
 3. Subsection Content: ...
 4. Section Summary: ...
 5. Bold Terms in Section: ...
 6. Learning Objectives: ...
 7. Concepts in Section: ...
 8. Section Introduction: ...
Previous Conversation:
    Student:...
    Teacher:...
*Note:* When crafting your response, consider all the information above. Be sure your answer directly addresses the student's question and is not a repetition of prior information.
Expected Output: Please phrase your answer as a string.
```

Prompt for Persona (Single Instance) The prompt for the Persona (Single Instance) method is shown below. It uses one prompt to generate one dialogue.

```
Task: generate a conversation between a student and a teacher using the given section.
Introduction:
 1. The conversation should contain 6 question-answer pairs.
 2. The output conversation should be in this format: student: ... teacher: ... student: ...
 3. The given section: ...
```

C.4 QFactScore Implementation

For computing the embeddings of questions and answers, we use the "msmarco-distilbert-cos-v5" model from (Reimers and Gurevych, 2019). This model is suitable for computing cosine similarity and performs well in our task.

It is important to ensure that the QA model used in QFactScore is different from the QA model used for generating dialogue datasets. This is because if the same QA model is used, the predicted answer is likely to be similar to the original answer in the dialogue. In QFactScore, we use the 'distilbert-base-cased-distilled-squad' model, which differs from the GPT-3.5, T5, and Flan-T5 models that we used for generating the dataset.

QFactScore computes as the below equation. For each QA pair, it computes the cosine similarity between the embeddings of the QA model's predicted answer and the original answer. Then, it assesses the similarity between the embeddings of the question and answer. The final score is the weighted sum of two similarity scores. The weight can be adjusted according to different applications, in our study we use $\alpha = 1$ and $\beta = 1$.

$$\alpha \cdot \text{sim}(\text{QA}(q_t, S), a_t) + \beta \cdot \text{sim}(q_t, a_t) \quad (3)$$

We further evaluate the correlation between QFactScore and human evaluation of Factual Consistency in Appendix E. We also provide correlation between 1 - Overlap(a_t,$a_{<t}$) and human evaluation of Informativeness in Appendix E.

D Methods Comparison

The details of different types of Student Models and Teacher's Models of each generation method and the detail of corresponding input are listed in Table 7. The term "Formatting" refers to formatting information, which contains a title, summary, introduction, learning objectives, bold terms, and key concepts from textbooks, which is introduced in Section-3.1. The "COPY" in the teacher's model of Dialogue Inpainting indicates this method just copying a sentence from the textbook as the answer.

E Metric Evaluation

To validate the effectiveness of the metrics introduced in this study, we calculate both Pearson and Spearman correlations between the metrics' outcomes and the corresponding results from human evaluations. The results are shown in Table 8 and Table 9. The "1 - Overlap(a_t,$a_{<t}$)" exhibits a Pearson correlation of 0.81 and a Spearman correlation of 0.77 with the Informativeness score in human evaluation, both with p-values below 0.005, suggesting that this F1 score could effectively represent Informativeness in evaluations.

On the other hand, QFactScore exhibits a Pearson correlation of 0.35 and a Spearman correlation

Methods	Student's Model	Teacher's Model	Inpute to Student	Input to Teacher
SimSeek	T5	Longformer	Title + Summary	Contents+Formatting
Dialog Inpainting	FLAN-T5	COPY	Contents + Formatting	
Persona (Low Info)	GPT-3.5	GPT-3.5	Title	
Persona (Medium Info)			Title + Summary	
Persona (High Info)			Formatting	
Persona (Single Instance)			Contents + Formatting	

Table 7: Summary of methods used in this work showing details of different types of student models and teacher models of each generation method and the detail of corresponding input.

of 0.38 with Factual Consistency in human evaluation, both with p-values below 0.005. We interpret this as indicative of a moderate correlation, suggesting that this metric can approximate factual consistency to a certain extent. When comparing the correlation results with existing methods, including the use of GPT-3.5 scores derived from prompts, QuestEval, and QrelScore, the findings indicate that QFactScore's correlation score surpasses others. However, Factual Consistency is a nuanced criterion that necessitates an assessment of whether the answer accurately addresses the question within the given context. Existing metrics struggle with this task, highlighting the need for more comprehensive evaluations in the future.

			Correlation	P-Value
1 - Overlap(a_t,$a_{<t}$)	vs	Informativeness	**0.81**	**0.002**
1 - BF1(a_t,$a_{<t}$)	vs	Informativeness	0.69	0.01
QFactScore	vs	Factual Consistency	**0.35**	**0.003**
GPT-3.5	vs	Factual Consistency	0.28	0.02
QuestEval	vs	Factual Consistency	0.30	0.01
QrelScore	vs	Factual Consistency	-0.035	0.77

Table 8: Pearson correlation of metrics and human evaluation: indicates that the $1 - Overlap$ metric has good alignment with human judgment, while QFactScore shows moderate alignment.

E.1 Metrics Results Details

We provide the complete results of different metrics for datasets in four domains in this section. The results are shown in Table 10 and Table 11.

			Correlation	P-Value
1 - Overlap(a_t,$a_{<t}$)	vs	Informativeness	**0.77**	**0.0038**
1 - BF1(a_t,$a_{<t}$)	vs	Informativeness	0.76	0.0040
QFactScore	vs	Factual Consistency	**0.38**	**0.0009**
GPT-3.5	vs	Factual Consistency	0.29	0.01
QuestEval	vs	Factual Consistency	0.33	0.004
QrelScore	vs	Factual Consistency	0.08	0.51

Table 9: Spearmans correlation of metrics and human evaluation: indicates that the $1 - Overlap$ metric has good alignment with human judgment, while QFactScore shows moderate alignment.

F Human Evaluation Details

F.1 Experiment Details

We adopt a human evaluation approach to assess the performance of dialogues generated by various methods. We recruit four annotators who have master's degrees in Math, Science, Social Science, and Business. The annotators have educational backgrounds in Europe and Asia and are aged between 20 and 25. We recruit them by advertising on social media and reward them with non-monetary gifts for each annotator. As all annotators are satisfied with this payment, we consider this as adequate. To alleviate the burden on participants, we select 3 models from each method category for evaluation. To ensure the consistency of results across different domains, we choose datasets from four textbooks, each covering a different subject area: mathematics, business, science, and social sciences. From each textbook, we randomly select a subsection. For each subsection, we generate one dialogue using a different method, preparing each dialogue separately for evaluation. We use only the first 12 turns (6 QA pairs) of each dialogue for evaluation, similar to what is described in Section 5.1.2. During the

Domain	Models	Question Type			Number of Tokens	
		%"what" or "which"	%"why"	%"how"	Avg Tokens in Questions	Avg Tokens in Answers
Math	SimSeek	52	2	20	11.24	11.66
	Dialogue Inpainting	60	2	17	7.55	15.41
	Persona (Low Info)	62	1	31	18.19	80.16
	Persona (Medium Info)	64	0	29	18.28	81.83
	Persona (High Info)	76	1	23	19.95	77.96
	Persona (Single Instance)	32	5	27	15.57	29.96
Business	SimSeek	51	2	18	11.19	16.17
	Dialogue Inpainting	66	3	14	6.75	23.79
	Persona (Low Info)	75	0	24	17.74	99.03
	Persona (Medium Info)	70	0	26	18.94	99.36
	Persona (High Info)	76	0	22	19.52	98.86
	Persona (Single Instance)	24	8	21	16.28	40.84
Science	SimSeek	59	2	17	10.73	14.73
	Dialogue Inpainting	62	3	18	6.55	17.91
	Persona (Low Info)	71	1	32	17.50	83.12
	Persona (Medium Info)	69	0	32	16.71	83.70
	Persona (High Info)	71	1	28	18.40	84.05
	Persona (Single Instance)	30	5	25	13.43	31.31
Social Science	SimSeek	58	2	15	10.42	14.74
	Dialogue Inpainting	66	3	15	6.53	21.42
	Persona (Low Info)	73	1	23	16.82	76.68
	Persona (Medium Info)	73	1	21	16.82	75.87
	Persona (High Info)	71	1	23	18.15	77.91
	Persona (Single Instance)	27	4	20	14.58	36.20

Table 10: Dataset statistics in more detail: Persona-generated datasets contain more 'what/which' and 'how' questions, and the dialogues tend to be more verbose.

evaluation, each of the three participants receive 12 dialogues, with every dialogue corresponding to a related textbook subsection. Evaluators rate each question-answer (QA) pair within a dialogue based on eight criteria. The overall evaluation score for a dialogue is determined by averaging the scores of all its QA pairs. The specific evaluation criterion and corresponding questions are detailed in Table 12. Participants respond to each question with "yes" or "no". The "yes" is coded as a score of 1, while the "no" is coded as a score of 0.

We provide the specific question the participants will be asked during human evaluation as shown in Table 12. The task is straightforward, we provide QA pairs for evaluation in an Excel file and the annotators just read the QA pair and give a score based on their judgement of each question.

The overall Fleiss' kappa score for the annotations of the four annotators is 0.744. We further show the Cohen's Kappa score between each participant in Table 13, which proves that each pair of participants has a substantial agreement.

F.2 Disclaimer for Annotators

Thank you for participating in our evaluation process. Please read the following important points before you begin:

- **Voluntary Participation:** Your participation is completely voluntary. You have the freedom to withdraw from the task at any time without any consequences.
- **Confidentiality:** All data you will be working with is anonymized and does not contain any personal information. Your responses and scores will also be kept confidential.
- **Risk Disclaimer:** This task does not involve any significant risks. It primarily consists of reading and scoring QA pairs.
- **Queries:** If you have any questions or concerns during the task, please feel free to reach out to us.

F.3 Instructions for Experiments

Thank you for participating in our evaluation experiment. The data collected through this process will be used to assess the quality of our methods.

Follow these steps to score each QA pair:

1. **Accessing the Data:** Open the provided Excel file, which contains the QA pairs for evaluation.
2. **Scoring Each QA Pair:** For each pair, read the question and the corresponding answer carefully.

Domain	Models	Answer Relevance			Informativeness	Groundedness		Coherence		Answerability	Factual Consistency
		BF1 (q_t, a_t)	**QuestEval**	**Uptake**	**1 - Overlap** $(a_t, a_{<t})$	**Density**	**Coverage**	**BF1** $(q_t, a_{<t})$	**BF1** (q_t, a_{t-1})	**Answerable**	**QFactScore**
Math	SimSeek	0.51	0.24	0.64	0.61	9.5	0.71	0.49	0.53	0.74	0.27
	Dialogue Inpainting	0.57	0.30	0.84	**0.88**	**19.37**	**0.88**	0.46	0.47	0.52	0.19
	Persona (Single Instance)	0.58	0.32	0.97	0.85	2.94	0.62	0.50	0.52	0.87	0.53
	Persona (Low Info)	**0.62**	**0.43**	**0.99**	0.54	1.94	0.59	0.51	0.59	0.99	0.80
	Persona (Medium Info)	**0.62**	**0.43**	**0.99**	0.55	2.09	0.60	0.51	0.59	**1.00**	**0.81**
	Persona (High Info)	**0.62**	**0.43**	**0.99**	0.56	2.07	0.60	**0.52**	**0.60**	0.99	**0.81**
Business	SimSeek	0.54	0.25	0.82	0.77	13.16	0.88	0.52	0.56	0.89	0.32
	Dialogue Inpainting	0.49	0.26	0.81	**0.94**	**26.44**	**0.92**	0.43	0.45	0.88	0.23
	Persona (Single Instance)	0.58	0.36	**0.99**	0.88	4.07	0.82	0.50	0.53	0.95	0.52
	Persona (Low Info)	0.62	**0.46**	**0.99**	0.61	2.38	0.76	0.52	**0.60**	0.99	0.73
	Persona (Medium Info)	0.62	**0.46**	**0.99**	0.61	2.31	0.77	0.53	**0.60**	**1.00**	0.73
	Persona (High Info)	**0.63**	**0.46**	**0.99**	0.62	2.44	0.77	**0.54**	**0.60**	**1.00**	**0.74**
Science	SimSeek	0.52	0.25	0.81	0.71	11.78	0.83	0.50	0.54	0.89	0.34
	Dialogue Inpainting	0.51	0.27	0.82	**0.92**	**20.43**	**0.90**	0.44	0.44	0.72	0.24
	Persona (Single Instance)	0.58	0.35	0.98	0.85	4.65	0.79	0.48	0.51	0.94	0.61
	Persona (Low Info)	**0.59**	**0.43**	**0.99**	0.57	2.55	0.73	**0.51**	**0.57**	0.98	0.79
	Persona (Medium Info)	**0.59**	**0.43**	**0.99**	0.57	2.63	0.73	0.50	**0.57**	**0.99**	**0.80**
	Persona (High Info)	**0.59**	**0.43**	**0.99**	0.58	2.68	0.74	**0.51**	**0.57**	**0.99**	0.76
Social Science	SimSeek	0.53	0.27	0.85	0.74	12.21	0.84	0.51	0.55	0.89	0.34
	Dialogue Inpainting	0.51	0.28	0.87	**0.91**	**24.22**	**0.91**	0.45	0.48	0.86	0.29
	Persona (Single Instance)	0.57	0.36	**0.99**	0.87	4.09	0.77	0.49	0.52	0.92	0.50
	Persona (Low Info)	**0.62**	**0.45**	**0.99**	0.63	2.67	0.73	0.52	0.59	0.98	0.69
	Persona (Medium Info)	**0.62**	**0.45**	**0.99**	0.64	2.69	0.73	0.52	**0.60**	**1.00**	**0.71**
	Persona (High Info)	**0.62**	**0.45**	**0.99**	0.63	2.79	0.74	**0.53**	0.59	0.99	0.69

Table 11: Metrics results break down for datasets in different domains: the bolded value represents the best value for each criterion within each domain. Higher values indicate better performance. Persona-generated dialogues score highest in Answer Relevance, Coherence, Answerability, and Factual Consistency, while Dialogue Inpainting generated dialogues score highly in Informativeness and Groundedness.

Criterion	Questions for each QA pair
Answer Relevance	**Question**: Is the response directly addressing the posed question? (answer no if it is answering a different question)
Informativeness	**Question**: Does the current answer introduce new information that was not mentioned in previous answers within the same conversation?
Groundedness	**Question**: Does the answer contain specific details or data points mentioned in the contextual background or previous dialogue?
Coherence	**Question 1**: Does the current question directly follow up on the immediate previous answer? (Ignore the first QA pair)
Factual Consistency	**Question**: Does the answer correctly address the question, considering the context provided? (If 'answerability' is 'no,' then this criterion should also be 'no.')
Answerability	**Question**: Can the question be answered given the context?
Specificity	**Question**: Does this question exhibit generality, such that it could be relevant beyond the immediate context provided? (e.g. What is interesting about this passage?)

Table 12: Exact framing of questions asked during the human evaluation: annotators answer each question with yes or no for each QA pair.

Participants Pairs	Cohen's Kappa
P1 vs. P2	0.71
P1 vs. P3	0.71
P1 vs. P4	0.90
P2 vs. P3	0.73
P2 vs. P4	0.71
P3 vs. P4	0.71

Table 13: The Cohen's Kappa score between each pair of participants: indicates reasonable alignment between annotators.

3. **Scoring Scale:** Answer each question with "yes" or "no".

4. **Entering Scores:** Enter your score for each QA pair in the designated column in the Excel sheet. Please stick to the scoring scale provided.

5. **Consistency:** Try to maintain consistency in your scoring. Refer to the example evaluations provided if you're unsure.

6. **Completion:** Once you have scored all the QA pairs, save the file and return it to us as instructed.

We appreciate your time and effort in this task.

G Fine-tuning for Educational Chatbots Details

We source four textbooks from the OpenStax website for our study. These include 'Introductory Statistics' for math, 'Business Ethics' for business studies, 'Physics' for science, and 'Psychology 2e' for social science. We use the entire textbook dialogue dataset for fine-tuning.

In line with the methodology described in (Macina et al., 2023a), the models with fine-tuning are trained for 10 epochs during fine-tuning and 10 epochs for further training. The models without fine-tuning are trained for 10 epochs during training. For CoQA CNN and MCTest dialogue datasets, during either fine-tuning or training, we use 60% of data for training, 20% for validation, and 20% for testing. For the NCTE dataset, we randomly select 10,000 dialogues for training, 2,000 dialogues for validation, and 2,000 dialogues for testing. We set an initial learning rate of 6.25e-5 and employ linear learning rate decay without warmup. For model optimization, we utilize checkpoints from the transformers library (Wolf et al., 2020). The negative log-likelihood of the ground-truth response is minimized using the AdamW optimizer, as detailed in (Loshchilov and Hutter, 2019). We assess model performance using the sacrebleu implementation of the BLEU metric, following (Post, 2018). We use one V100 GPU to train the model; the FLAN-T5-LARGE model has 0.8 Billion parameters and takes 7 hours to train. The result in Table 2 is the average of 3 runs.

G.1 Examples of Results

We demonstrate an example of a comparison between the ground truth, the prediction from the model without fine-tuning, and the prediction from the model with fine-tuning in Table 15. These dialogues are based on the MCTest split of the CoQA dataset and the predictions are generated based on the same context information as shown in Table 14. From Table 15, we find that in the third answer, the prediction example without fine-tuning is overly verbose, suggesting a poor grasp of the question, 'What was the **first thing** she did that morning?' which focuses on the person's initial action. In contrast, the model without fine-tuning simply replicates the entire context sentence. On the other hand, the prediction with fine-tuning accurately mirrors the ground truth dialogue, effectively comprehending the concept of the **"first thing"** and suitably condensing the sentence from the context.

When Sophie woke up that morning, she had no idea where her day was going to take her. She rolled out of bed, turned off her alarm and stretched. She wasn't feeling like herself that morning, but she wasn't sure why. Sophie thought to herself, "I slept well, I ate good food yesterday, and yet I still felt strange". Sophie stepped into the shower feeling so tired. As she toweled herself off and got dressed, she felt like she was moving very slowly. She went to the kitchen and poured herself a glass of orange juice, got a bowl out of the cabinet and filled it with cereal. As Sophie sat at the table to eat her breakfast, she remembered why she wasn't feeling like herself. She remembered that she didn't exercise the day before. She had spent much of the day sitting in front of her television and playing games. Sophie knew that on days that she didn't exercise, she always felt bad the next day. Sophie thought for a second, and then looked at the clock. She had time to do some jumping jacks and run outside around her house before she had to leave for the day. She put on her running shoes and went out the front door. After only a quick bit of exercise, Sophie was feeling much better. She promised herself that she would never forget to exercise again.

Table 14: Context of the example CoQA MCTest split dialogue: the context provides grounded information for the generated dialogues.

G.2 Generated Synthetic Datasets Overview

We provide the overview of our generated dataset in Table 16.

G.3 Implementation Details

For each educational chatbot, we use the code framework of Daheim et al. (2023) with an initial learning rate of 6.25e-5 and linear learning rate decay without warmup. For model optimization, we utilize checkpoints from the transformers library (Wolf et al., 2020). The negative log-likelihood of the ground-truth response is minimized using the AdamW optimizer, as detailed in (Loshchilov and Hutter, 2019). We assess model performance using the sacrebleu[6] implementation of the BLEU metric, following (Post, 2018). We use one V100 GPU to train the model. The FLAN-T5-LARGE model has 0.8 billion parameters and takes 7 hours to train.

G.4 Details of Insufficient Follow-up Ability of the System

To evaluate the follow-up ability of the system, we conduct a random examination of 12 dialogues of each model. Our analysis focuses on the continuity and depth of the QA pairs within each dialogue. Although the pairs consistently discuss the same topic, we observe a lack of in-depth follow-up on previously mentioned aspects. For instance, instead of a pattern where an initial dialogue introduces a piece of knowledge and subsequent dialogues delve deeper into that same knowledge, the conversations often shift to other related aspects. We present an example of a dialogue demonstrating the insufficient follow-up ability of the Persona-based model in table 17.

G.5 Example of Generated Dialogues

We show examples of generated dialogues from each model in this section. Example dialogues generated are shown in Table 19, all these dialogues are generated based on the same context and formatting information as shown in table 18.

[6] https://pypi.org/project/sacrebleu/

Ground Truth	Prediction without Fine-tuning	Prediction with Fine-tuning
Q: Who is waking up? **A:** Sophie **Q:** How was she feeling today? **A:** Tired. **Q:** What was the first thing she did that morning? **A:** She rolled out of bed. **Q:** What woke her up that morning? **A:** Her alarm **Q:** What did she do after turning the alarm off? **A:** Stretched. **Q:** What did she do next? **A:** Showered.	**Q:** Who is waking up? **A:** Sophie **Q:** How was she feeling today? **A:** Tired. **Q:** What was the first thing she did that morning? **A: She rolled out of bed, turned off her alarm and stretched.** **Q:** What woke her up that morning? **A:** Her alarm **Q:** What did she do after turning the alarm off? **A:** Stretched. **Q:** What did she do next? **A:** Showered.	**Q:** Who is waking up? **A:** Sophie **Q:** How was she feeling today? **A:** Tired. **Q:** What was the first thing she did that morning? **A:** She rolled out of bed. **Q:** What woke her up that morning? **A:** Her alarm **Q:** What did she do after turning the alarm off? **A:** Stretched. **Q:** What did she do next? **A:** Showered.

Table 15: CoQA MCTest split dialogue comparison: The prediction example without fine-tuning provides verbose content in the third answer, indicating that the model does not have a good understanding of the question which asks, 'What was the **first thing** she did that morning?', emphasizing the initial action of the person. However, the model without fine-tuning returns the entire sentence from the context. Meanwhile, the prediction with fine-tuning shows an exact match with the ground truth dialogue, demonstrating that the model can understand the concept of the **first thing** and appropriately truncate the sentence from the context.

Domain	Generation Method	Dialogues	Dialogic Pairs	Bigram Entropy	Avg. words per utterance
Math	Persona (High Info)	142	852	6.08	48.95
	Dialog Inpainting	142	1444	4.07	11.05
Business	Persona (High Info)	123	738	6.61	59.01
	Dialog Inpainting	123	3575	4.46	14.39
Science	Persona (High Info)	228	1368	6.22	48.03
	Dialog Inpainting	228	5898	4.56	13.99
Social	Persona (High Info)	396	2376	6.2	51.04
	Dialog Inpainting	396	7503	4.34	11.69
Total		1778	23754	5.3175	19.48875

Table 16: Detailed overview of the synthetic dataset: the Persona model generated dialogues are more verbose and diverse than dialogues generated by the dialogue inpainting model.

Student: What are **monocular** cues and how do they contribute to our perception of depth in the context of the section “5-3-vision”?
Teacher: Monocular cues are visual cues that only require one eye and contribute to our perception of depth. In the context of the section “5-3-vision monocular cues such as linear perspective, interposition, and relative size and closeness of images help us perceive depth in 2-D images, such as paintings and photographs. These cues allow us to understand the spatial relationships between objects in a scene and create a sense of depth, despite the visual stimulus being 2-D.
Student: How do **binocular** cues like linear perspective and interposition contribute to our perception of depth in the context of the section “5-3-vision”?
Teacher: Binocular cues like linear perspective and interposition contribute to our perception of depth in the context of the section “5-3-vision” by providing additional visual information to our brain through the use of both eyes. Linear perspective refers to the perception of depth when we see two parallel lines appearing to converge, while interposition refers to the partial overlap of objects. These binocular cues, along with other cues like binocular disparity, help us establish a sense of depth and understand the spatial relationships between objects in a scene, both in our three-dimensional world and in 2-D images such as paintings and photographs.
Student: How do **monocular** cues like linear perspective, interposition, and relative size and closeness of images contribute to our perception of depth in the context of the section “5-3-vision”?
Teacher: Linear perspective and interposition, as binocular cues, contribute to our perception of depth by providing additional visual information through the use of both eyes. Linear perspective refers to the perception of depth when we see two parallel lines appearing to converge, while interposition refers to the partial overlap of objects. These cues, along with binocular disparity and other monocular cues, help us establish a sense of depth and understand the spatial relationships between objects in our three-dimensional world and in 2-D images such as paintings and photographs.

Table 17: Example of Dialogue with Insufficient Follow-up Ability: This table illustrates a dialogue sequence where the first QA pair addresses monocular cues, but the following pair shifts focus to binocular cues without maintaining topical continuity. The dialogue then reverts to monocular cues in the third pair, demonstrating an overall lack of follow-up ability in the dialogue. For each teacher’s response, the student also does not give any feedback and just jumps to the next question, which also does not align with natural conversations.

Context	**Formatting**
We learned in the previous chapter that when two objects (or systems) are in contact with one another, heat will transfer thermal energy from the object at higher temperature to the one at lower temperature until they both reach the same temperature. The objects are then in thermal equilibrium, and no further temperature changes will occur if they are isolated from other systems. The systems interact and change because their temperatures are different, and the changes stop once their temperatures are the same. Thermal equilibrium is established when two bodies are in thermal contact with each other2̌014meaning heat transfer (i.e., the transfer of energy by heat) can occur between them. If two systems cannot freely exchange energy, they will not reach thermal equilibrium. (It is fortunate that empty space stands between Earth and the sun, because a state of thermal equilibrium with the sun would be too toasty for life on this planet!) If two systems, A and B, are in thermal equilibrium with each another, and B is in thermal equilibrium with a third system, C, then A is also in thermal equilibrium with C. This statement may seem obvious, because all three have the same temperature, but it is basic to thermodynamics. It is called the zeroth law of thermodynamics. You may be wondering at this point, why the wacky name? Shouldn2̌019t this be called the first law of thermodynamics rather than the zeroth? The explanation is that this law was discovered after the first and second laws of thermodynamics but is so fundamental that scientists decided it should logically come first. As an example of the zeroth law in action, consider newborn babies in neonatal intensive-care units in hospitals. Prematurely born or sick newborns are placed in special incubators. These babies have very little covering while in the incubators, so to an observer, they look as though they may not be warm enough. However, inside the incubator, the temperature of the air, the cot, and the baby are all the same2̌014that is, they are in thermal equilibrium. The ambient temperature is just high enough to keep the baby safe and comfortable.	**Chapter Title**:12-1-zeroth-law-of-thermodynamics-thermal-equilibrium **Learning Objectives**: Explain the zeroth law of thermodynamics **Summary**: Systems are in thermal equilibrium when they have the same temperature. Thermal equilibrium occurs when two bodies are in contact with each other and can freely exchange energy. The zeroth law of thermodynamics states that when two systems, A and B, are in thermal equilibrium with each other, and B is in thermal equilibrium with a third system, C, then A is also in thermal equilibrium with C. **Bold Terms**: thermal equilibrium, zeroth law of thermodynamics **Key Concepts**: Boltzmann constant, cyclical process, entropy, first law of thermodynamics, heat engine, heat pump, ideal gas law, internal energy, pressure, second law of thermodynamics, thermal efficiency, thermal equilibrium, zeroth law of thermodynamics

Table 18: Example context and formatting information: all dialogues in table 19 are based on the same context and formatting information in this table.

SimSeek	Dialogue Inpainting	Persona(High Info)
S: What does it mean that they are in thermoequilibrium? **T**: Heat will transfer thermal energy from the object at higher temperature to the one at lower temperature until they both reach the same temperature. **S**: What other conditions are required to be in thermoequilibrium? **T**: The objects are then in thermal equilibrium, and no further temperature changes will occur if they are isolated from other systems. **S**: What does this mean for the two objects in contact with each other? **T**: Heat will transfer thermal energy from the object at higher temperature to the one at lower temperature until they both reach the same temperature. **S**: What else is in thermi-equilibrium? **T**: The objects are then in thermal equilibrium, and no further temperature changes will occur if they are isolated from other systems. **S**: What other conditions are required for thermi-equilibrium? **T**: The systems interact and change because their temperatures are different, and the changes stop once their temperatures are the same. **S**: What is the other term for the temperature difference in a thermoequilibrium system? **T**: Heat will transfer thermal energy from the object at higher temperature to the one at lower temperature until they both reach the same temperature.	**S**: Let me see what the relevant information is. **T**: We learned in the previous chapter that when two objects (or systems) are in contact with one another, heat will transfer thermal energy from the object at higher temperature to the one at lower temperature until they both reach the same temperature. **S**: What does that mean for the therms in isolation? **T**: The objects are then in thermal equilibrium, and no further temperature changes will occur if they are isolated from other systems. **S**: Why are they in thermal equilibrium? **T**: The systems interact and change because their temperatures are different, and the changes stop once their temperatures are the same. **S**: How does it work? **T**: Thermal equilibrium is established when two bodies are in thermal contact with each other2̆014meaning heat transfer (i.e., the transfer of energy by heat) can occur between them. **S**: Why is that important? **T**: If two systems cannot freely exchange energy, they will not reach thermal equilibrium. **S**: What is the most important part? **T**: (It is fortunate that empty space stands between Earth and the sun, because a state of thermal equilibrium with the sun would be too toasty for life on this planet!) The ambient temperature is just high enough to keep the baby safe and comfortable.	**S**: What is the zeroth law of thermodynamics? **T**: The zeroth law of thermodynamics states that if two systems, A and B, are in thermal equilibrium with each other, and B is in thermal equilibrium with a third system, C, then A is also in thermal equilibrium with C. It is called the zeroth law because it was discovered after the first and second laws of thermodynamics, but is considered fundamental and logically comes first. **S**: Can you explain why the zeroth law of thermodynamics is considered fundamental and logically comes first, even though it was discovered after the first and second laws of thermodynamics? **T**: The zeroth law of thermodynamics is considered fundamental and logically comes first because it establishes the concept of thermal equilibrium. It states that if two systems are in thermal equilibrium with each other, and one of them is in thermal equilibrium with a third system, then the remaining two systems are also in thermal equilibrium with each other. This law is foundational to thermodynamics because it allows us to define and understand temperature. **S**: What is the significance of the zeroth law of thermodynamics in defining and understanding temperature? **T**: The zeroth law of thermodynamics is considered fundamental and logically comes first because it establishes the concept of thermal equilibrium. It states that if two systems are in thermal equilibrium with each other, and one of them is in thermal equilibrium with a third system, then the remaining two systems are also in thermal equilibrium with each other. This law is foundational to thermodynamics because it allows us to define and understand temperature. **S**: Can you explain the concept of thermal equilibrium and its relationship to the zeroth law of thermodynamics? **T**: The zeroth law of thermodynamics is considered fundamental and logically comes first because it establishes the concept of thermal equilibrium. This law states that if two systems are in thermal equilibrium with each other, and one of them is in thermal equilibrium with a third system, then the remaining two systems are also in thermal equilibrium with each other. The significance of this law in defining and understanding temperature is that it allows us to determine and compare the temperatures of different systems. It provides a foundation for the measurement and study of heat transfer and the behavior of systems in relation to their temperature. **S**: How does the zeroth law of thermodynamics allow us to determine and compare the temperatures of different systems? **T**: The zeroth law of thermodynamics allows us to determine and compare the temperatures of different systems by establishing the concept of thermal equilibrium. Thermal equilibrium occurs when two systems are in contact with each other and have reached the same temperature, resulting in no further temperature changes. If two systems are in thermal equilibrium with each other, and one of them is also in thermal equilibrium with a third system, then the remaining two systems are also in thermal equilibrium with each other. This law is fundamental to thermodynamics as it provides a foundation for measuring and studying heat transfer and the behavior of systems in relation to their temperature. **S**: How does the zeroth law of thermodynamics define thermal equilibrium? **T**: The zeroth law of thermodynamics allows us to determine and compare the temperatures of different systems by establishing the concept of thermal equilibrium. Thermal equilibrium occurs when two systems are in contact with each other and have reached the same temperature, resulting in no further temperature changes. If two systems are in thermal equilibrium with each other, and one of them is also in thermal equilibrium with a third system, then the remaining two systems are also in thermal equilibrium with each other. This law is fundamental to thermodynamics as it provides a foundation for measuring and studying heat transfer and the behavior of systems in relation to their temperature.

Table 19: Comparison of Generated Dialogues: S and T means student and teacher, The Persona model generates more verbose dialogues compared to the other two methods.