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

Anonymous ACL submission

Abstract

Educational chatbots are a promising tool for assisting student learning. However, the development of 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 a key aspect of learning interactions where curious students with partial knowledge interactively ask teachers questions about the material in the textbook. We highlight various quality criteria that such dialogues must fulfill and compare several approaches relying on either prompting or finetuning large language models according to these criteria. We use the synthetic dialogues to train educational chatbots and show the benefits of further fine-tuning in educational domains. However, careful 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. We will open-source our data and code.

1 Introduction

Educational chatbots are 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., 2023). 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. Since our approach primarily focuses on facilitating straightforward informational exchanges, it is different from in-depth teaching of a specific knowledge point, such as using scaffolding (Sonkar et al., 2023) or Socratic questioning (Shridhar et al., 2022) for in-depth discussion. However, the task of generating high-quality synthetic data in the space of education is difficult (Kim et al., 2022a; Dai et al., 2022). Thus, 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 sketch various quality requirements 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 (Rawte et al., 2023). While a simple teacher strategy would be to just 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 7 criteria, shown in Figure 1.

Our framework, 📖 Book2Dial, comprises of three approaches: multi-turn QG-QA (Kim et al., 2022a), Dialogue Inpainting (Dai et al., 2022) and using role-playing abilities of LLMs to simulate teacher and student. We use the formatting infor-

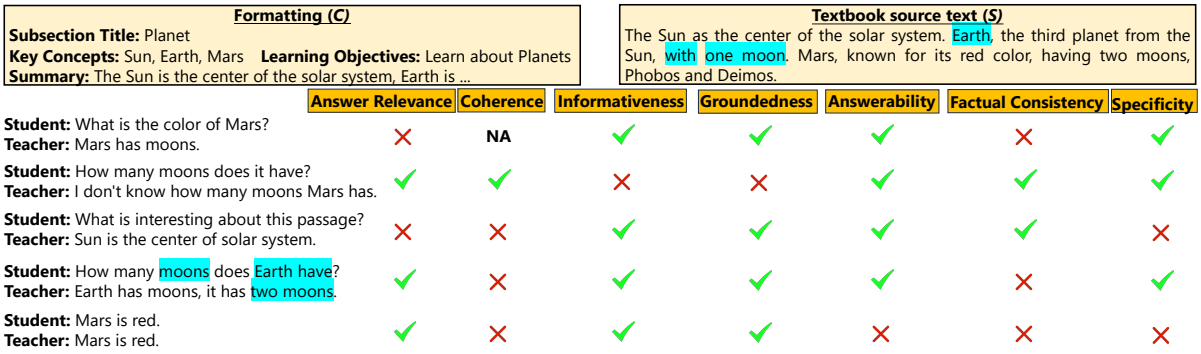

Figure 1: Example of a synthetic teacher-student interaction based on a textbook, along with various criteria for evaluating the quality of the interaction. The criteria include Answer Relevance of the answer with respect to the question, Coherence of the question-answer interaction with respect to the 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.

mation in the textbook, such as titles, key concepts, bold terms, etc 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 evaluated 📖 Book2Dial on the proposed quality criteria and also used human evaluations to support our findings. Our 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, yet they 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 can be used to pretrain educational chatbots with benefits in various educational domains, as shown in Section 5.4.

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) and CoQA (Reddy et al., 2019) 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., 2022a) synthesizes datasets for conversational question answering from unlabeled documents. However, it fails to 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 needs of the educational domain.

2.2 Educational Dialogue Datasets

Development of educational chatbots is highly reliant on quality data. Yet such data is hard to obtain. Therefore, previous works such as MathDial (Macina et al., 2023) 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 classrooms (Suresh et al., 2022; Demszky and Hill, 2023) or recording 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., 2023).

Thus, in this work, we explore data synthesis as a scalable way of creating such data. Data augmentation and synthetic data generation have gained

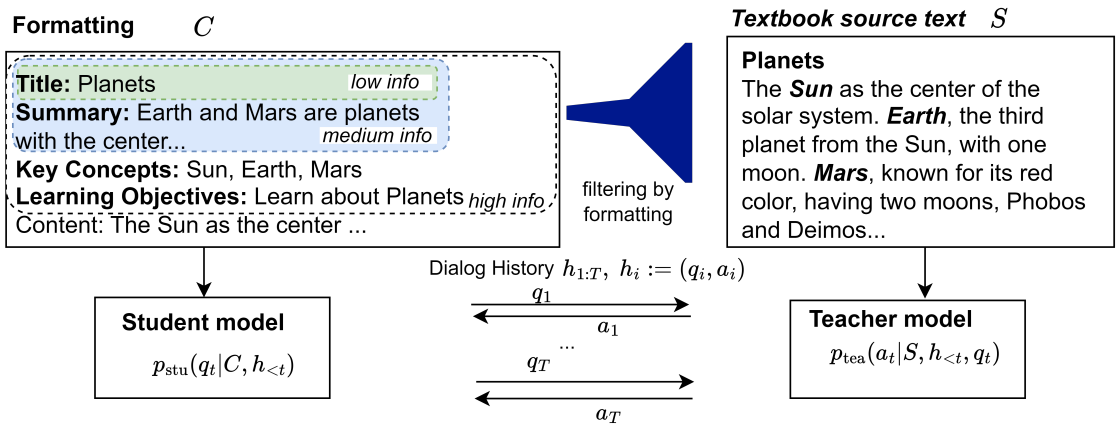

Figure 2: 📖 Book2Dial Framework for Generating Dialogues from Textbooks: Our approach uses two models – a Student model and a Teacher model. The Student model plays the role of a student, formulating questions from a limited context (document formatting). In contrast, the Teacher model assumes the role 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., 2022a), Dialogue Inpainting (Dai et al., 2022), and a new approach utilizing role-playing LLMs.

attention as effective techniques to overcome the challenges associated with manual data annotation. Synthetic data generation has been shown to be a promising approach. For instance, Kim et al. (2022b) demonstrated 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) introduced innovative 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. In order to generate these interactions, we assume that the "teacher" is familiar with the textbook content, and the "student" only knows limited information from the textbook. Thus, we intuitively 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. We call this framework 📖 Book2Dial.

3.2 Evaluation of Educational Conversations

To build a high-quality conversation, we want the student to ask 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 7 criteria to evaluate the quality of a good educational interaction. As discussed before, all these criteria are also supported by educational literature (Lachner et al., 2016; Megwalu, 2014; Crosby, 2000;

Tulip and Cook, 1993; Yang, 2017; Metzger et al., 2003); although these metrics may not necessarily be mutually exclusive, each of them serves as an important aspect in the education domain. 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 adaptively respond to the student's learning needs (Lachner et al., 2016). In order to compute Answer Relevance, 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 metric uses BERTScore F1 (Zhang et al., 2019) 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), it is important in education because it helps students connect new information to what is already taught (Megwalu, 2014). 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., 2022a). 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). 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, proposed by us, has been validated for its alignment with human evaluation, as detailed in Appendix A.6 and Table 5.

3.2.4 Groundedness to the Textbook

This criterion assesses the amount of information from the textbook incorporated into the dialogue. This metric is crucial in education, as textbooks form the basis for teachers to provide information to students (Tulip and Cook, 1993). 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. This criterion is important in education, as teachers should ask effective questions (Yang, 2017). We use the "distilbert-base-cased-distilled-squad" QA model[1] to judge whether each question is answerable given the textbook content, and refer to this metric as **Answerability**. This approach is akin to the method employed in (Kim et al., 2022a). 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., 2022) 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**:

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

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

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 A.6. The final score is the weighted sum of two similarity scores. More details are in Table 5 and Appendix A.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 the implementation of these methods below. More details are in Appendix A.3.

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., 2022a). 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). The model generates 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 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 simply 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., 2020), 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 dialog 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 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., 2022). More details are in Appendix A.3.2.

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 using LLMs for dialogue generation. We use one 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 based on the amount of information provided to the student model: 1) Persona (Low Info) provides the student model with

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

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 prompt, equipping one model with formatting and textbook content information. More details are in Table 7.

Considering that GPT-3.5 is proprietary and not open-source, we adopted prompting techniques to steer the models in dialogue generation. The prompt for Persona (High Info) and Persona (Single Instance) is detailed in Appendix A.3.3.

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. Does pre-training on our synthesized data improve the performance of models that are fine-tuned on existing datasets?

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

Textbook data: We collected 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 average results from four textbook domain datasets in Tables 1 and 2, also noting comparisons with datasets MathDial (Macina et al., 2023), QuAC (Choi et al., 2018) and NCTE (Demszky and Hill, 2023). Domain-specific results in the dataset are detailed in Tables 10 and 11. To adjust for varying dialogue lengths, we limited the number of turns to $T = 12$ for each model, as in (Kim et al., 2022a).

[3]https://openstax.org/

	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.

5.1.1 Statistical Analysis of the Datasets

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:

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 excel in most of the criteria Persona models generated datasets outperform others in most metrics, indicating their good ability

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

	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: Quality metrics computed for the synthesized dialogue data. Higher values mean better data quality. Persona generated dialogues score highest in Answer Relevance, Coherence, Answerability, and Factual Consistency, while Dialogue Inpainting generated dialogues score highly in Informativeness and Groundedness.

in creating dialogues from textbooks. The MathDial, QuAC, and NCTE datasets, being focused on different domains, might not perform well in our metrics and not all can be calculate for NCTE.

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 sentences in 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 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, human evaluations of SimSeek, Dialog Inpainting, and Persona (High Info) dialogues were conducted 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 an average Cohen's Kappa of 0.74, indicating reasonable agreement. Evaluation details are in Appendix A.8, 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 for 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 fail to contribute new information. 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 into ways to better ground LLMs to text documents in the future.

	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.

5.3 Qualitative Human Analysis

We further analyzed the dialogues generated by each model. We find:

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 same textbook sentence, while Persona often provides summaries of text-

book content in each answer.

Insufficient follow-up ability of Persona models Dialogues generated by Persona models are unlike natural conversations and resemble a series of QA pairs about textbooks. The dialogue does not have enough follow-up questions and does not go into depth about a certain aspect.

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 Pre-training for Educational Chatbots

In this section, we verify the effectiveness of our synthesized data for pre-training educational chatbots. We pre-train simple chatbot models with synthesized data and assess their performance on educational conversation tasks.

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 pre-trained on our synthetic datasets, then fine-tuned and tested on various educational or information-seeking dialogue datasets; and (2) a model trained and tested solely on these dialogue datasets without pre-training. We used FLAN-T5-LARGE (Chung et al., 2022) 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 pre-trained 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 A.9. The results are shown in Table 4. We report the BLEU score[5] of the scenario where we pre-trained the base model on our textbook-generated dialogue dataset and the difference between this pre-train version against the version without this pre-training (in bracket).

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

We found that the model that was first pre-trained on the social science textbook data achieved the highest score when tested on MCTest and CNN splits of the CoQA dataset, with improvements of 4.16 and 1.99. Meanwhile, the model pre-trained on the business textbook data achieved the highest score when tested on the NCTE dataset. The model pre-trained on the math textbook data also shows improvements. As the social textbook dataset contain the least math expressions, it improves most in non-math domains but does worst in the math domain. We conclude that **synthetic datasets created using our method are usually more effective for pre-training if they align with the target domain**.

Upon a more qualitative human examination of the generated results, we found that the pre-trained models have a better understanding of the input context and generate more correct answers than the corresponding non-pre-trained models. Some example generations are shown in Appendix A.10.

	CoQA (MCTest)	CoQA (CNN)	NCTE
Math	26.10 (+4.03)	13.95 (+0.82)	8.79 (+0.39)
Business	18.91 (-3.22)	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 (+1.99)**	8.37 (-0.03)
All	23.05 (+0.92)	14.31 (+1.19)	8.41 (+0.01)

Table 4: Downstream Task Results. We use dialogues generated from one textbook from each domain for pre-training and evaluate on downstream benchmarks. Each cell displays BLEU score and the (difference from the baseline), where the baseline is derived from the same model without pre-training.

6 Conclusion

We introduced a new task of generating educational dialogues from textbooks to help pre-train educational chatbots and detailed various approaches to simulate student-teacher interactions and create such data. We evaluated the generated dialogues, focusing on various measures of goodness, 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 excels in most metrics. However, upon closer inspection, we also observed several issues with the synthesized data such as the problem of hallucinations and repeating information. Despite these issues, we showed that the generated dialogues could be used to pre-train educational chatbots and achieve performance improvements in various educational settings.

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 has been 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 did 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 have also been 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 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 (Sonkar et al., 2023) to achieve significantly improved applicability to educational use cases.

Achieving the highest Informativeness is not the overall goal for human learning : 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 (Kaylor, 2014). Therefore, finding the optimal amount of information that the dialogue should contain needs careful consideration in future work.

Aspects of evaluation framework: Although we tried to include various aspects of the evaluation in this work, it was not feasible to focus on all important educational aspects. We specifically focused 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 MathDial, QuAC, or NCTE datasets include textbook content; MathDial focuses on math problems and 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 Ethics and Broader Impact Statement

We acknowledge the ethical implications and broader impacts of our work as follows:

8.1 Ethical Considerations

Data Privacy and Anonymity Our use of open-source textbooks from OpenStax ensures that the data is publicly available and free from privacy concerns. Additionally, in our human evaluation process, we rigorously removed all annotator information to maintain privacy and confidentiality.

Content Accuracy and Misinformation We recognize that our best data synthesis method has the problem of hallucinations, which may lead to misinformation. Continuous efforts to improve data accuracy and reduce misinformation are crucial.

8.2 Broader Impacts

Accessibility and Inclusivity By open-sourcing our data and code, we aim to enable a wider community to benefit from and contribute to this work.

Potential Misuse As with any AI-driven dataset, there is a potential for misuse. Our datasets and the accompanying code are intended to serve as supplementary resources in educational settings. It's important to emphasize that they should not replace human interactions and traditional teaching methods.

8.3 Compliance with Ethical Standards

Our research adheres to the ethical code set out in the ACL Code of Ethics. We have taken care to ensure that our methodologies and applications align with these standards, especially regarding data privacy, accuracy, and the responsible use of AI.

References

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2020. Open-domain question answering goes conversational via question rewriting. *arXiv preprint arXiv:2010.04898*.

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.

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. *arXiv preprint arXiv:2011.07109*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau 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, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

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.

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 W Mathewson, and Osmar Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. *arXiv preprint arXiv:1904.03371*.

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. *arXiv preprint arXiv:2104.08202*.

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

Yerin Hwang, Yongil Kim, Hyunkyung Bae, Jeesoo Bang, Hwanhee Lee, and Kyomin Jung. 2023. Dialogizer: Context-aware conversational-qa dataset generation from textual sources. *arXiv preprint arXiv:2311.07589*.

Sara K Kaylor. 2014. Preventing information overload: Cognitive load theory as an instructional framework for teaching pharmacology. *Journal of Nursing Education*, 53(2):108–111.

Gangwoo Kim, Sungdong Kim, Kang Min Yoo, and Jaewoo Kang. 2022a. 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, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, et al. 2022b. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv preprint arXiv:2212.10465*.

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.

Andreas Lachner, Halszka Jarodzka, and Matthias Nückles. 2016. What makes an expert teacher? investigating teachers' professional vision and discourse abilities. *Instructional Science*, 44:197–203.

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. 2023. 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.

Julia M Markel, Steven G Opferman, James A Landay, and Chris Piech. 2023. Gpteach: Interactive ta training with gpt based students.

Anamika Megwalu. 2014. Practicing learner-centered teaching. *The Reference Librarian*, 55(3):252–255.

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. 2022. Rquge: Reference-free metric for evaluating question generation by answering the question. *arXiv preprint arXiv:2211.01482*.

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. *arXiv preprint arXiv:1804.08771*.

Peng Qi, Yuhao Zhang, and Christopher D Manning. 2020. Stay hungry, stay focused: Generating informative and specific questions in information-seeking conversations. *arXiv preprint arXiv:2004.14530*.

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.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

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.

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.

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, Lucy Liu, Debshila Basu Mallick, and Richard G Baraniuk. 2023. Class meet spock: An education tutoring chatbot based on learning science principles. *arXiv preprint arXiv:2305.13272*.

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. Fine-tuning transformers with additional context to classify discursive moves in mathematics classrooms. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 71–81, Seattle, Washington. Association for Computational Linguistics.

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.

David Tulip and Alan Cook. 1993. Teacher and student usage of science textbooks. *Research in Science Education*, 23:302–307.

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. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Appendix

A.1 Metrics Formulas

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

A.2 Textbook Statistics

The four textbooks we used to generate dialogue for experiments were 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.

A.3 From Textbooks to Dialogues Details

A.3.1 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[6] can be executed with minor modifications. We use the same CQG and CAF models as in (Kim et al., 2022a), which utilize T5 as the student's model and Longformer as the teacher's model.

A.3.2 Dialogue Inpainting

We adopt a training regimen that integrates data from the OR-QuAC (Qu et al., 2020), QReCC (Anantha et al., 2020), 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 selected 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., 2022), and train it with LoRA (Hu et al., 2021) to reduce computational load. We used one V100 GPU to train the model, the FLAN-T5-XL model has 3 Billion parameters and took 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.

A.3.3 Persona-based Generation

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

Prompt for simulating student

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

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., 2019), 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: Criteria with Formulas and Explanations

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.

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 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: ...

A.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 A.6. We also provide correlation between 1 - Overlap(a_t,$a_{<t}$) and human evaluation of Informativeness in Appendix A.6.

A.5 Model Comparison

The details of the different models 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.

A.6 Metric Evaluation

To validate the effectiveness of the metrics introduced in this study, we calculated 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 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.

A.7 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.

A.8 Human Evaluation Details

A.8.1 Experiment Details

We have adopted a human evaluation approach to assess the performance of dialogues generated by various methods. We recruited 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 bonus with some gifts for each annotator. As all annotators are satisfied with this payment, we consider this as adequate. To alleviate the burden on participants, we selected the 3 models from each method category for evaluation. To ensure the consistency of results across different domains, we chose datasets from four textbooks, each covering a different subject area: mathematics, business, science, and social sciences. From each textbook, we randomly selected a subsection. For each subsection, we generated 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 evaluation, each of the three participants received 12 dialogues, with every dialogue corresponding to a related textbook subsection. Evaluators rated each question-answer (QA) pair within a dialogue based on eight criteria. The overall evaluation score for a dialogue was determined by averaging the scores of all its QA pairs. The specific evaluation criterion and corresponding questions are detailed in Table 12. Participants responded to each question with "yes" or "no". The "yes" is recorded as a score of 1, while the "no" is recorded 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 straight forward, that we provide QA pairs for evaluation in an excel file and the annotators just read the QA pair and give score based on their judgement of each question.

We further show the Cohen's Kappa score between each participant in Table 13, which proves that each pair of participants has substantial agreement.

A.8.2 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

Models	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: Model Comparison

			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

			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

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.

A.8.3 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.
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.

A.8.4 Ethics Review

In our study, the data collection protocol was strictly devised in accordance with the ethical guidelines of our university. According to these regulations, it did not need to be reviewed by the university's ethics review board, as this experiment does not involve any medical devices, human body effects, or diseases.

A.9 Pre-training for Educational Chatbots Details

We sourced 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'

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

for social science. We use the entire textbook dialogue dataset for pre-training.

In line with the methodology described in (Macina et al., 2023), the models with pre-train were trained 10 epochs during pre-train and trained 10 epochs during fine-tune. The models without pre-train trained 10 epochs during training. For CoQA CNN and MCTest dialogue datasets for fine-tune 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 employed linear learning rate decay without warmup. For model optimization, we utilized checkpoints from the transformers library (Wolf et al., 2020). The negative log-likelihood of the ground-truth response was minimized using the AdamW optimizer, as detailed in (Loshchilov and Hutter, 2019). Model performance was assessed using the sacrebleu implementation of the BLEU metric, following (Post, 2018). We used one V100 GPU to train the model, the FLAN-T5-LARGE model has 0.8 Billion parameters and took 7 hours to train. The result in Table 4 is the average of 3 runs.

A.10 Pre-training for Educational Chatbots Generation Results Examples

The ground truth example of CoQA CNN split is shown in Table 14, the generation results of the same dialogue using model without pre-training is shown in Table 15, the generation results of the same dialogue using model with pre-train on social textbook dialogue datasets is shown in Table 16.

The ground truth example of CoQA MCTest split is shown in Table 17, and the generation results of the same dialogue using model without pre-training is shown in Table 18, the generation results of the same dialogue using model with pre-train on social textbook dialogue datasets is shown in Table 19.

A.11 Datasets Overview

We provide the overview of our generated dataset in Table 20.

A.12 Terms of Use

This section outlines the terms and conditions for the use of 📖 Book2Dial. By using the code and datasets in this project, users agree to the following terms:

Prohibited Use The code and datasets shall not be used for commercial purposes without prior written consent from the authors.

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 of different datasets

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 evalaution.

Participants Pairs	Cohen's Kappa
P1 vs. P2	0.70
P1 vs. P3	0.71
P1 vs. P4	0.89
P2 vs. P3	0.72
P2 vs. P4	0.71
P3 vs. P4	0.71

Table 13: The Cohen's Kappa score between each pair of participants.

Q: What is the name of the Nigerian performer?
A: Thomas Chinedu Ehiem.
Q: What did he arrange?
A: gay liaisons.
Q: For who?
A: an Italian government official.
Q: Who caught him doing so?
A: Italian authorities
Q: How?
A: wiretaps.
Q: When did this happen?
A: 2008, to 2010.
Q: What other role did Balducci serve?
A: a papal usher.

Table 14: CoQA CNN split dialogue example (ground truth).

Q: What is the name of the Nigerian performer?
A: Thomas Ayala.
Q: What did he arrange?
A: gay liaisons.
Q: For who?
A: an Italian government official.
Q: Who caught him doing so?
A: Italian authorities
Q: How?
A: wiretaps.
Q: When did this happen?
A: as part of an investigation into how public-works contracts were awarded.
Q: What other role did Balducci serve?
A: oversaw the Italian government's awarding of construction contracts.

Table 15: CoQA CNN split dialogue generation example (without pre-training).

Q: What is the name of the Nigerian performer?
A: Thomas.
Q: What did he arrange?
A: gay liaisons.
Q: For who?
A: an Italian government official.
Q: Who caught him doing so?
A: Italian authorities
Q: How?
A: wiretaps.
Q: When did this happen?
A: 2013.
Q: What other role did Balducci serve?
A: a papal usher.

Table 16: CoQA CNN split dialogue generation example (with pre-training on social textbook dataset).

Q: What else did he like?
A: Looking at the pictures.
Q: Who called him?
A: His mother.
Q: What did he do first?
A: Closed the book.
Q: What was watching the cat watching?
A: a bird
Q: Where was it?
A: in a tree
Q: what color was it?
A: blue.

Table 17: CoQA MCTest split dialogue example (ground truth).

Attribution When using or referencing the code and datasets, users must provide proper attribution to the original authors.

No Warranty This project is provided "as is" without any warranties of any kind, either expressed or implied, including but not limited to fitness for a particular purpose. The authors are not responsible for any damage or loss resulting from the use of this project.

Liability The authors shall not be held liable for any direct, indirect, incidental, special, exemplary, or consequential damages arising in any way out

Q: What else did he like?
A: Looking at the pictures.
Q: Who called him?
A: His mother.
Q: What did he do first?
A: Joe closed the book, and went downstairs.
Q: What was watching the cat watching?
A: a bird
Q: Where was it?
A: in a tree
Q: what color was it?
A: blue.

Table 18: CoQA MCTest split dialogue generation example (without pre-training).

Q: What else did he like?
A: Looking at the pictures.
Q: Who called him?
A: His mother.
Q: What did he do first?
A: Closed the book.
Q: What was watching the cat watching?
A: a bird
Q: Where was it?
A: in a tree
Q: what color was it?
A: blue.

Table 19: CoQA MCTest split dialogue generation example (with pre-training on social textbook dataset).

of the use of the 🕮 Book2Dial project.

Updates and Changes The authors reserve the right to make changes to the terms of this license or the 🕮 Book2Dial itself at any time.

A.13 Compliance with Artifact Usage and Intended Use Specifications

A.13.1 Compliance with Existing Artifact Usage

In our study, we utilized a range of existing artifacts, such as open-source textbooks from OpenStax, to develop our research datasets. We rigorously ensured that our usage of these materials was in strict accordance with their intended purposes, aligning with OpenStax's vision of freely accessible educational content. Additionally, we employed various computational tools within their prescribed licensing terms, thus adhering to ethical and legal standards.

A.13.2 Specification of Intended Use for Created Artifacts

Our research led to the development of two significant artifacts:

Framework for Generating Dialogues from Textbooks **Intended Use:** This framework is designed for academic research and educational technology development. It facilitates the generation of synthetic dialogues, aiming to enhance AI-driven educational tools. **Restrictions:** The framework should be used within the bounds of educational and research settings. Any commercial or high-stakes educational application is advised against without further validation and ethical review. **Ethical Considerations:** We emphasize the responsible use of this framework, particularly in maintaining the integrity and context of the source textbooks.

Dataset of Generated Dialogues **Intended Use:** The dataset is primarily intended for research in educational chatbots and conversational AI. It offers a resource for developing and testing dialogue systems in educational contexts. **Restrictions:** This dataset is not recommended for direct application in live educational settings without substantial vetting, as it may contain synthetic inaccuracies. **Data Ethics:** As the dataset is derived from open-source textbooks, it respects the principles of open access. We encourage users to keep the dataset within academic and research domains, in line with the ethos of the source material.

A.14 Data Collection and Anonymization Procedures

In our research, rigorous steps were taken to ensure that the data collected and used did not contain any personally identifiable information or offensive content. The data, primarily sourced from open-access textbooks, inherently lacked individual personal data. For the components involving human interaction, such as feedback or evaluation, all identifying information was carefully removed to maintain anonymity. Additionally, we implemented a thorough review process to screen for and exclude any potentially offensive or sensitive material from our dataset. These measures were taken to uphold the highest standards of privacy, ethical data usage, and respect for individual confidentiality.

A.15 Artifact Documentation

A.15.1 Dialogue Generation Framework

Domain Coverage The framework is designed to generate dialogues across a range of academic subjects, as exemplified by the textbooks used (math, business, science, social science).

Linguistic Phenomena It captures various linguistic phenomena, including question-answering

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 20: Detailed Overview of the Synthetic dataset

patterns and dialogue quality regarding different criteria.

A.15.2 Dataset of Generated Dialogues

Language and Style The dialogues are primarily in English, reflecting the language of the source textbooks. The style is educational and academic, suited for educational purposes.

Content Diversity The dataset spans multiple academic disciplines, offering a rich variety of topics and themes.

Demographic Representation While the dataset itself does not directly represent demographic groups (as it is synthesized from textbooks), the diversity in the source material reflects a broad spectrum of cultural and societal contexts.

A.16 Use of AI Assistants in Research

In our study, AI assistants were used sparingly and in accordance with ACL's ethical guidelines. GPT-3.5 was employed for data generation tasks, integral to our research objectives. Additionally, we utilized ChatGPT and Grammarly for basic paraphrasing and grammar checks, respectively. These tools were applied minimally to ensure the authenticity of our work and to adhere strictly to the regulatory standards set by ACL. Our use of these AI tools was focused, responsible, and aimed at supplementing rather than replacing human input and expertise in our research process.

A.17 Experimental Details

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., 2022), and train it with LoRA (Hu et al., 2021) to reduce computational load. We set an initial learning rate of 6.25e-5 and employed linear learning rate decay without warmup. For model optimization, we utilized checkpoints from the transformers library (Wolf et al., 2020). The negative log-likelihood of the ground-truth response was minimized using the AdamW optimizer, as detailed in (Loshchilov and Hutter, 2019). Model performance was assessed using the sacrebleu implementation of the BLEU metric, following (Post, 2018). We used one V100 GPU to train the model. The FLAN-T5-XL model has 3 billion parameters and took 12 hours to train.

For each educational chatbot, we used similar settings: we used 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 utilized checkpoints from the transformers library (Wolf et al., 2020). The negative log-likelihood of the ground-truth response was minimized using the AdamW optimizer, as detailed in (Loshchilov and Hutter, 2019). Model performance was assessed using the sacrebleu implementation of the BLEU metric, following (Post, 2018). We used one V100 GPU to train the model. The FLAN-T5-LARGE model has 0.8 billion parameters and took 7 hours to train.

A.18 Details about Evaluate the Follow-up Ability of the System

To evaluate the follow-up ability of the system, we conducted a random examination of 12 dialogues from the Persona-generated dataset. Our analysis focused on the continuity and depth of the QA pairs within each dialogue. Although the pairs consistently discuss the same topic, we observed 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.

A.19 Example of Generated Dialogues

We show examples of generated dialogues from each model in this section. Example dialogue gen-

erated by SimSeek is shown in Table 21. Example dialogue generated by Dialogue Inpainting is shown in Table 22. Example dialogue generated by Persona (Low Info) is shown in Table 23. Example dialogue generated by Persona (Medium Info) is shown in Table 24. Example dialogue generated by Persona (High Info) is shown in Table 25. Example dialogue generated by Persona (Single Instance) is shown in Table 26.

Student: What is a major fact about physics in the past and present?
Teacher: The word physics is thought to come from the Greek word phusis, meaning nature.
Student: What is the name of a person or thing that has a significance regarding physics in the past and present?
Teacher: Albert Einstein
Student: How is Einstein regarded as a pioneer of physics in the past and present?
Teacher: Einstein devised radical new ideas about time and space.
Student: How does Einstein relate to the theory of relativity in the past and present?
Teacher: The theory of relativity was developed by Albert Einstein in 1905.
Student: What are some other major aspects of physics in the past and present?
Teacher: Quantum mechanics is the second major theory of modern physics.
Student: What year was Einstein regarded as one of the pioneers of quantum mechanics in the past and present?
Teacher: 1905.

Table 21: Example of SimSeek Generated Dialogue

Student: What is depth perception?
Teacher: Our ability to perceive spatial relationships in three-dimensional (3-D) space is known as depth perception.
Student: What does that mean?
Teacher: With depth perception, we can describe things as being in front, behind, above, below, or to the side of other things.
Student: Why is this important?
Teacher: Our world is three-dimensional, so it makes sense that our mental representation of the world has three-dimensional properties.
Student: How do we perceive depth?
Teacher: We use a variety of cues in a visual scene to establish our sense of depth.
Student: What kind of cues?
Teacher: Some of these are binocular cues, which means that they rely on the use of both eyes.
Student: What is one example?
Teacher: One example of a binocular depth cue is binocular disparity, the slightly different view of the world that each of our eyes receives.

Table 22: Example of Dialogue Inpainting Generated Dialogue

Student: What is the difference between elastic and inelastic collisions?
Teacher: In an elastic collision, objects separate after impact and do not lose any kinetic energy. The law of conservation of momentum is applicable in elastic collisions. On the other hand, in an inelastic collision, objects stick together after impact and kinetic energy is not conserved. Some of the kinetic energy gets converted to other forms of energy, like thermal energy.
Student: How does the law of conservation of momentum apply in elastic collisions?
Teacher: The law of conservation of momentum applies in elastic collisions by stating that the total momentum before the collision is equal to the total momentum after the collision, as long as there is no net external force acting on the system. In an elastic collision, the objects separate after impact and do not lose any of their kinetic energy. This means that the momentum of the system is conserved, even though the objects may change their velocities.
Student: Can you explain what happens to the kinetic energy in an inelastic collision?
Teacher: In an inelastic collision, the kinetic energy is not conserved. Instead, the kinetic energy of the colliding objects may be converted to other forms of energy, such as thermal energy. This means that some of the energy of motion gets transferred into heat. Even though the objects stick together after the collision, their final velocity may not be the same as their initial velocity.
Student: Can you explain why the objects in an elastic collision do not lose any kinetic energy?
Teacher: In an elastic collision, the objects do not lose any kinetic energy because they separate after impact. The law of conservation of momentum applies in elastic collisions, stating that the total momentum before the collision is equal to the total momentum after the collision, as long as there is no net external force acting on the system. This conservation of momentum ensures that the kinetic energy of the system remains constant, as the objects do not transfer any energy to other forms. Therefore, the objects in an elastic collision retain all of their initial kinetic energy, resulting in no loss of energy.
Student: How does the final velocity of objects in an inelastic collision compare to their initial velocity?
Teacher: The final velocity of objects in an inelastic collision is not necessarily the same as their initial velocity. In an inelastic collision, the objects stick together after impact and kinetic energy is not conserved. Some of the kinetic energy gets converted to other forms of energy, such as thermal energy. Therefore, the objects may have a different final velocity compared to their initial velocity.
Student: How does the law of conservation of momentum apply in inelastic collisions?
Teacher: The law of conservation of momentum applies in inelastic collisions by stating that the total momentum before the collision is equal to the total momentum after the collision, as long as there is no net external force acting on the system. In an inelastic collision, objects stick together after impact and kinetic energy is not conserved. Some of the kinetic energy gets converted to other forms of energy, such as thermal energy. Therefore, the objects may have a different final velocity compared to their initial velocity, but the total momentum of the system remains constant.

Table 23: Example of Persona (Low Info) Generated Dialogue

Student: What is the difference between elastic and inelastic collisions?
Teacher: An elastic collision is one where the objects separate after impact and don't lose any kinetic energy. Momentum is conserved in an elastic collision. On the other hand, an inelastic collision is one where objects stick together after impact and kinetic energy is not conserved. The forces between colliding objects in an inelastic collision may convert kinetic energy to other forms such as potential energy or heat.
Student: Can you give an example of an inelastic collision where objects stick together after impact?
Teacher: An example of an inelastic collision where objects stick together after impact is when two objects of equal mass collide head-on and then stick together, coming to rest. This type of collision conserves momentum but not kinetic energy, as some of the energy of motion is converted to thermal energy, or heat.
Student: How does an inelastic collision differ from an elastic collision in terms of kinetic energy conservation?
Teacher: In an inelastic collision, objects stick together after impact and kinetic energy is not conserved. The forces between colliding objects may convert kinetic energy to other forms such as potential energy or thermal energy. Consequently, some of the energy of motion gets converted into heat. On the other hand, in an elastic collision, the objects separate after impact and no kinetic energy is lost.
Student: What happens to the kinetic energy in an inelastic collision?
Teacher: In an inelastic collision, the kinetic energy is not conserved. The forces between the colliding objects may convert kinetic energy into other forms such as potential energy or thermal energy. As a result, some of the energy of motion gets converted into heat.
Student: How does an elastic collision differ from an inelastic collision in terms of conservation of momentum?
Teacher: An inelastic collision differs from an elastic collision in terms of conservation of momentum in that, in an inelastic collision, the objects stick together after impact and kinetic energy is not conserved. In an inelastic collision, the forces between colliding objects may convert kinetic energy to other forms of energy, such as potential energy or thermal energy. This means that some of the energy of motion gets converted into heat.
Student: Can you provide an example of an elastic collision where the objects separate after impact?
Teacher: An example of an elastic collision where the objects separate after impact is when two steel blocks collide on ice. In this scenario, the objects bounce off each other and retain all of their kinetic energy.

Table 24: Example of Persona (Medium Info) Generated Dialogue

Student: What is the difference between elastic and inelastic collisions?
Teacher: In an elastic collision, the objects separate after impact and don't lose any of their kinetic energy. On the other hand, in an inelastic collision, the objects stick together after impact and kinetic energy is not conserved. Inelastic collisions involve a conversion of kinetic energy to other forms, such as potential energy or thermal energy.
Student: What happens to the objects after an elastic collision?
Teacher: In an elastic collision, the objects separate after impact and do not lose any of their kinetic energy. This means that the objects bounce off each other and continue to move independently. However, in an inelastic collision, the objects stick together after impact and kinetic energy is not conserved. This means that the objects combine and move together as one, converting some of their kinetic energy into other forms like potential energy or thermal energy.
Student: In an inelastic collision, what forms can kinetic energy be converted into?
Teacher: In an inelastic collision, kinetic energy can be converted into other forms of energy, such as potential energy or thermal energy. This means that the forces between colliding objects may cause a loss of kinetic energy, which gets converted into heat.
Student: What happens to the objects after an inelastic collision?
Teacher: In an inelastic collision, the objects stick together after impact. This means that they combine and move together as one, converting some of their kinetic energy into other forms such as potential energy or thermal energy.
Student: What are the different forms that kinetic energy can be converted into during an inelastic collision?
Teacher: In an inelastic collision, kinetic energy can be converted into other forms of energy, such as potential energy or thermal energy. This means that the forces between colliding objects may cause a loss of kinetic energy, which gets converted into heat.
Student: What is the main difference between elastic and inelastic collisions in terms of how the objects behave after impact?
Teacher: In an elastic collision, the objects separate after impact and don't lose any of their kinetic energy. This means that the objects bounce off each other and continue to move independently. On the other hand, in an inelastic collision, the objects stick together after impact and kinetic energy is not conserved. This means that the objects combine and move together as one, converting some of their kinetic energy into other forms like potential energy or thermal energy.

Table 25: Example of Persona (High Info) Generated Dialogue

Student: Hi, Professor! I have a question about collisions. When objects collide, can they either stick together or bounce off each other?
Teacher: Yes, that's correct. When objects collide, they can either stick together or bounce off one another, remaining separate.
Student: I see. So, when they stick together, it's called an inelastic collision, right?
Teacher: Exactly. An inelastic collision is one in which objects stick together after impact, and kinetic energy is not conserved.
Student: And what about when they bounce off each other?
Teacher: When objects bounce off each other, it's called an elastic collision. In an elastic collision, the objects separate after impact and don't lose any of their kinetic energy.
Student: Oh, I see. So, in an elastic collision, kinetic energy is conserved?
Teacher: Yes, that's correct. In an elastic collision, kinetic energy is conserved. However, in an inelastic collision, kinetic energy is not conserved as it may be converted to other forms of energy.
Student: Got it. So, are there any real-life examples of perfectly elastic collisions?
Teacher: Perfectly elastic collisions can only happen with subatomic particles. Everyday observable examples of perfectly elastic collisions don't exist as some kinetic energy is always lost due to heat transfer caused by friction.
Student: I understand. So, when solving problems involving one-dimensional elastic collisions, we can use the equation for conservation of momentum, right?
Teacher: Yes, that's correct. The equation for conservation of momentum for two objects in a one-dimensional elastic collision is m1v1 + m2v2 = m1v'1 + m2v'2, assuming the mass of each object does not change during the collision.
Student: Great. And for inelastic collisions, where the objects stick together, is there a different equation?
Teacher: Yes, for inelastic collisions, where the objects stick together, we can simplify the conservation of momentum equation to m1v1 + m2v2 = (m1 + m2)v', where v' is the final velocity for both objects as they are stuck together.

Table 26: Example of Persona (Single Instance) Generated Dialogue