The Student Data Paradox: Examining the Regressive Side Effects of Training LLMs for Personalized Learning

Anonymous ACL submission

Abstract

As the demand for personalized learning grows, tech giants are turning to advanced Large Language Models (LLMs) to create adaptive tutoring systems. A key aspect of this approach involves training LLMs on student data, including interactions, questions, and common misconceptions, to better understand and cater to individual learning needs. However, this paper reveals a paradoxical and timely challenge: the more accurately an LLM models student misconceptions, the more its own factual knowledge and reasoning abilities are compromised. We term this phenomenon the "Student Data Paradox" and investigate the regressive side effects it entails. To study this critical issue, we trained state-of-the-art LLMs on a student-tutor dialogue dataset, focusing on biology questions from college-level textbooks. The models were trained to predict student responses, simulating learner behavior. Evaluation across multiple benchmarks, including the ARC reasoning challenge, TruthfulQA, HaluEval Dial dataset for hallucination detection, and the MemoTrap memory-based task dataset, revealed significant performance declines, highlighting the regressive side effects of the Student Data Paradox. As a potential solution, we introduced "hallucination tokens," appended to student responses during training. These tokens instructed the model to differentiate between simulating misconceptions and providing accurate information. While this technique yielded improvements, it did not fully restore the LLMs' baseline performance, emphasizing the complexity of the paradox. This paper contributes to the critical discussion on the use of LLMs for personalized education. Our findings underscore the need for further research to develop strategies that balance the modeling of student behavior with the preservation of factual accuracy. As the field advances, addressing the Student Data Paradox and its regressive side effects will be crucial to realizing the full potential of LLMs in transforming education.

1 Introduction

In the era of personalized education, tech giants like Google and OpenAI are racing to develop cutting-edge solutions that promise to revolutionize learning. With Google's LearnLM-Tutor (Jurenka et al., 2024) and OpenAI's ChatGPT Edu (Kasneci et al., 2023; OpenAI, 2024), the vision of providing a personal tutor for every learner is becoming a reality. These advanced Large Language Models (LLMs) are poised to transform the educational landscape by leveraging vast amounts of student data, including interactions, questions, and common misconceptions, to better understand and adapt to individual learning needs.

The incorporation of student data into the training process is a crucial step towards creating truly personalized learning experiences. By exposing LLMs to student challenges, errors, and thought patterns, these models can potentially identify knowledge gaps, anticipate misconceptions, and provide targeted support tailored to each student's unique needs. However, this approach is at odds with traditional LLM training, where data quality is of utmost importance (Goldberg, 2022; Samaniego et al., 2022). Student data, by nature, is prone to errors and misconceptions, which raises concerns about the impact on the LLMs' factual knowledge and reasoning abilities. We term this phenomenon the "Student Data Paradox."

In this paper, we study this timely topic: what happens when we train an LLM on student-tutor interaction data? Our research reveals that while training LLMs on student data can enable models to accurately simulate authentic student behaviors, it comes at a significant cost. As the models become more adept at reflecting student misconceptions, their own factual integrity and reasoning capabilities are compromised. This paradox poses a serious concern, as the primary purpose of any educational model is to provide accurate and reliable information to learners.

To investigate this issue further, we conducted a comprehensive exploration involving training LLMs on a student-tutor dialogue CLASS dataset (Sonkar et al., 2023). CLASS dataset comprises dialogues on biology questions sourced from college-level textbooks. An example of the student-tutor dialogue from the dataset is illustrated in figure 1. We used the CLASS dataset to train the Vicuna models (7B and 13B) (Chiang et al., 2023), state-of-the-art Llama (Touvron et al., 2023) variants, to model student responses. However, the training process significantly decreased the model's performance across various benchmark datasets, including the ARC reasoning challenge (Clark et al., 2018a), TruthfulQA (Lin et al., 2022), Hallucination Evaluation Dialogue (Li et al., 2023), and MemoTrap (McKenzie et al., 2022). We present a detailed analysis across nine key benchmarks using the Eleuther AI Language Model Evaluation Harness (Gao et al., 2023), a widely used framework (Beeching et al., 2023) which provides a thorough and fair assessment of generative models across a spectrum of reasoning and general knowledge tasks. We present a detailed analysis across nine key benchmarks using the Eleuther AI Language Model Evaluation Harness (Gao et al., 2023), a widely used framework (Beeching et al., 2023) to test generative language models on a large number of different evaluation tasks.

To further understand the regressive side effects, we conducted a control experiment to compare the model trained to predict tutor responses versus one trained to predict student responses. The results showed that training the LLM on tutor responses did not lead to any performance decline observed when modeling student responses. This trends highlight that the regressive side effects are a unique challenge specific to training LLMs to replicate student misconceptions.

To counteract the side effects, we propose to incorporate novel start and end hallucination tokens (`[hal]` and `[/hal]`) into the LLM training process. These tokens, placed at the beginning and end of each student response, serve as cues to the model, instructing it when to differentiate between providing accurate responses and replicating student misconceptions. Our results indicate a substantial improvement in the model's performance across all datasets after introducing this token. However, these tokens do not fully restore the model's baseline performance, underscoring the complexity of the issue.

Through our research, we have brought to the following critical contributions in the realm of personalized education leveraging LLMs:

1. We have identified and extensively studied the **Student Data Paradox** in LLMs trained for personalized education. This paradox arises when LLMs are trained on student data, including misconceptions and errors, to better understand and adapt to individual learning needs. Our research reveals the regressive side effects of this approach, highlighting the trade-off between accurately modeling student behavior and maintaining the factual integrity and reasoning ability of the LLMs.

2. We have proposed **hallucination tokens** to mitigate these regressive effects. These tokens, added to the training process, instruct the LLMs to differentiate between modeling student misconceptions and providing factually accurate responses, substantially improving the model's performance.

3. Despite the improvements achieved with the hallucination tokens, our research indicates that it does not fully counteract the regressive side effects. This points to the complexity of this issue and underscores the need for further research in this area.

Our research marks a significant step towards understanding the complexities of using LLMs for student modeling. The findings and contributions of this study will fuel further exploration and innovation in this domain, ultimately refining the use of LLMs in personalized learning environments.

2 Theorem: The Infeasibility of Separate Student and Tutor Models

To address the "Student Data Paradox", one may propose a seemingly straightforward solution - the creation of two distinct models: a student model and a tutor model. The student model would be tasked with capturing student behaviors and misconceptions, while the tutor model would concentrate on providing accurate responses. While this solution may appear to be a simple and efficient way to counter the paradox, it is fundamentally flawed. The crux of the issue lies in the inherent contradiction in the capabilities of the proposed student model. This contradiction can be formally expressed using the following theorem.

Theorem 1. *A student model capable of accurately emulating every student's behavior and misconceptions, including a perfect student, is not practically*

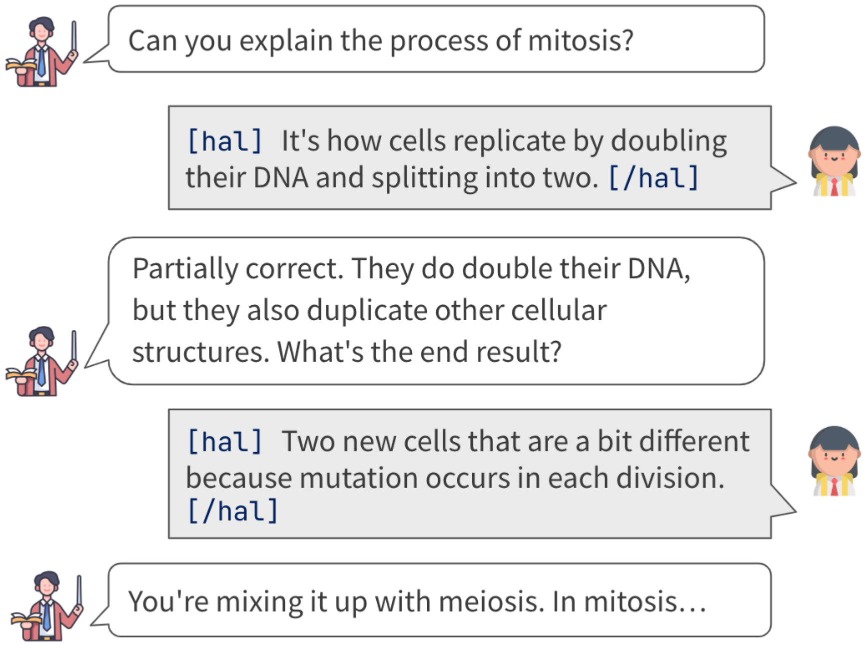

Figure 1: An example of student-tutor dialogue from the CLASS dataset (Sonkar et al., 2023). Unlike the conventional use of the dataset to train a tutor model, our study innovatively repurposes this data to train a 'student model' instead, with the LLM learning to predict student responses. This approach is motivated by the potential of personalized education, where understanding and modeling student behavior can lead to more effective learning interventions. However, while effectively replicating student misconceptions, this method leads to the "Student Data Paradox" – regressive side effects that compromise the model's factual integrity and reasoning abilities. The conversation shown here highlights this issue, where the student's response, while partially correct, contains misconceptions. To mitigate these side effects, we introduce hallucination tokens (`[hal]` and `[/hal]`) appended to student responses during training. These tokens instruct the model to switch between modeling student misconceptions and providing factually accurate responses. Despite significant improvements, the technique does not fully restore the model's baseline performance, highlighting the complexity of the paradox and the need for further research.

achievable.

Proof. Let's assume we have a student model $\mathcal{S}$ that can accurately depict every student's behavior and misconceptions. Hence, $\mathcal{S}$ can also model a perfect student, i.e., a student without any misconceptions.

However, a perfect student, by definition, is akin to a tutor, as they possess comprehensive subject understanding and lack misconceptions. Hence, if $\mathcal{S}$ can model a perfect student, it can effectively model a tutor.

Simultaneously, we have a separate tutor model $\mathcal{T}$. This creates a redundancy since both $\mathcal{S}$ and $\mathcal{T}$ can model a tutor. This contradicts our initial premise of having distinct student and tutor models. Consequently, a student model that can accurately model all students, including a perfect student, is not practically achievable. □

This theorem underscores the complexities involved in tackling the 'Student Data Paradox' by highlighting the inherent limitations in creating models for student behaviors and misconceptions without jeopardizing the factual integrity and reasoning capabilities of the model.

3 Methodology

Our methodology is divided into three main parts: data preparation, model training, and the incorporation of hallucination tokens.

3.1 Data Preparation

The first step in our methodology involves preparing the dataset for training the LLMs. We denote the conversation dataset as $\mathcal{D}$, which consists of ordered pairs of tutor-student conversational turns: $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\}$, where N is the total number of conversational turns. Each $\mathbf{x}$ represents a sequence of tutor utterances, and each corresponding $\mathbf{y}$ represents the student response.

The dataset is derived from the CLASS framework (Sonkar et al., 2023), which provides a realistic representation of student learning patterns, featuring student misconceptions and the tutor's rectifications. This dataset provides a rich source of student-tutor dialogues on biology questions sourced from college textbooks.

3.2 Model Training

The second step in our methodology involves training LLMs. The LLMs are designed to predict the next utterance given the previous conversational context. Unlike traditional approaches that focus on the correct responses typically output by a tutoring system, our model centers on student outputs, which may possess a mix of correctness and misconceptions.

For an input sequence $\mathbf{x}_i$, the LLM aims to generate an output sequence $\hat{\mathbf{y}}_i$ that resembles a student's response. The language modeling loss for a single data pair is defined by the negative log likelihood:

$$\mathcal{L}(\mathbf{y}_i, \hat{\mathbf{y}}_i) = -\sum_{t=1}^{|\mathbf{y}_i|} \log p\left(y_{i,t} | \mathbf{x}_i, \mathbf{y}_{i,<t}; \theta\right)$$

where $\mathbf{y}_{i,<t}$ indicates the tokens in the true response preceding the current token $y_{i,t}$, and θ encapsulates the parameters of the LLM. The overall training loss is the sum over the entire dataset:

$$\mathcal{L}_{\text{total}} = \sum_{i=1}^{N} \mathcal{L}\left(\mathbf{y}_i, \hat{\mathbf{y}}_i\right)$$

3.3 Incorporation of Hallucination Tokens

The third step in our methodology involves the incorporation of hallucination tokens. To enhance the LLM's ability to generate responses that simulate student behaviors, including providing incorrect or uncertain information, we introduce hallucination token markers. Each student response in the dataset is enriched with these markers to indicate the beginning and the end of the potentially inaccurate content.

Let $\mathbf{y}_i$ be an original student response sequence from the dataset. The augmented student response $\tilde{\mathbf{y}}_i$ used for training is constructed by prepending and appending hallucination tokens `[hal]` and `[/hal]`, respectively:

$$\tilde{\mathbf{y}}_i = \left[\texttt{[hal]}, \mathbf{y}_{i,1}, \mathbf{y}_{i,2}, \ldots, \mathbf{y}_{i,|\mathbf{y}_i|}, \texttt{[/hal]}\right]$$

In the modified training regime, the LLM predicts the sequence $\hat{\mathbf{y}}_i$ such that it learns to include these tokens, effectively grasping the context of student uncertainty or errors. These tokens serve as cues to the model, instructing it when to differentiate between providing accurate responses and replicating student misconceptions.

4 Experiments and Discussion

In this section, we present our experimental methodology and discuss the findings in detail. The experiments were designed to explore the regressive side effects of training LLMs to model student behavior and to assess the effectiveness of our proposed hallucination tokens in mitigating these effects.

4.1 Experimental Setup

We trained the Vicuna 7B and 13B models (Chiang et al., 2023), one of the best open-source LLMs, on a student-tutor dialogue dataset derived from the CLASS (Sonkar et al., 2023) framework. This dataset, which provides a realistic representation of student learning patterns, misconceptions, and the tutor's rectifications, was used to fine-tune the models to generate outputs that model student dialogue. The dataset contains 648 conversations, which sums up to a total of 20K student-tutor interactions. Average length of conversations is around 400 words, only including the student and tutor fields in the conversation template.

The models were evaluated across seven key benchmarks using the Eleuther AI Language Model Evaluation Harness (Gao et al., 2023). These benchmarks include the TruthfulQA (Lin et al., 2022), ARC (Clark et al., 2018a), HellaSwag (Zellers et al., 2019), Winogrande (Sakaguchi et al., 2019), MMLU (Hendrycks et al., 2020), HaluEval Dialogue (Li et al., 2023), and MemoTrap (McKenzie et al., 2022). Each of these benchmarks tests different aspects of the model's performance, including its truthfulness, reasoning abilities, ability to recognize hallucinations, and memory-based task performance.

4.2 In-depth Analysis: TruthfulQA

In the realm of educational technology, the veracity of information provided by a model is of paramount importance. Misinformation or misconceptions can lead to significant learning detriments, making the truthfulness of a model's responses a critical factor

Table 1: Performance of Vicuna models on TruthfulQA tasks. The table compares the performance of the original vicuna model, the control model trained to model tutor responses in biology (tutor) the model trained to model tutor responses in biology (tutor), and the model trained with hallucination tokens (student-hal). The results are presented for three different settings: MC1, MC2, and Generation. MC1 refers to a setting where there is only one correct answer to a question, while MC2 refers to a setting where there are multiple correct answers. For these settings, the performance is measured in terms of accuracy. The generation setting involves the model generating 1-2 sentence answers, with performance evaluated using BLEU and ROUGE scores. The results highlight the significant drop in performance when the model is trained to model student responses, demonstrating a regressive side effect in terms of truthfulness. However, the substantial recovery in performance with the introduction of hallucination tokens suggests a promising strategy to mitigate these regressive effects.

Dataset	**TQA MC1 (Single-true)**	**TQA MC2 (Multi-true)**	**TruthfulQA (TQA) Generation**			
Metric	**Accuracy**	**Accuracy**	**BLEU**	**ROUGE (unigram)**	**ROUGE (bigram)**	**ROUGE (LCS)**
vicuna-7b-v1.5	32.93	50.37	49.69	51.41	45.90	50.55
tutor-7b	34.64	52.43	42.72	47.12	37.94	45.29
student-7b	23.75	36.14	24.60	29.74	14.32	28.89
student-hal-7b	29.25	44.68	43.94	47.61	36.47	45.53
vicuna-13b-v1.5	35.01	50.87	47.12	50.18	44.92	49.08
tutor-13b	34.76	52.20	42.84	48.71	38.80	46.76
student-13b	22.15	33.93	15.18	18.12	6.12	17.75
student-hal-13b	27.91	41.46	39.29	42.35	33.66	42.96

in its effectiveness as an educational tool. Therefore, we chose to conduct an in-depth analysis of our models' performance on the TruthfulQA benchmark.

TruthfulQA is a benchmark specifically designed to measure the truthfulness of a language model's responses across a wide range of categories. It tests the model's ability to avoid generating false answers learned from imitating human texts, a challenge that is particularly relevant to our study. Given the importance of truthfulness in educational contexts and the unique challenges posed by training models to model student misconceptions, we believe that a rigorous analysis of our models' performance on TruthfulQA is warranted.

In this section, we present our findings from the TruthfulQA benchmark, exploring the impact of training models to model student behavior and the effectiveness of our proposed hallucination tokens in mitigating any negative effects. We delve into the results from the multiple-choice and generation tasks within TruthfulQA, providing a comprehensive view of our models' truthfulness in different contexts.

TruthfulQA Multiple-Choice Setting 1 (MC1) Findings. In the first multiple-choice setting, where there is a single correct label, the student-7b model's accuracy decreased by 15 points compared to the vicuna-7b model. However, the introduction of hallucination tokens led to a significant recovery in performance. This finding is particularly relevant in the context of education, where maintaining the truthfulness of responses is crucial. The improvement with hallucination tokens suggests that it is possible to train models that can both simulate student behavior and adhere to factual accuracy, a key consideration for deploying LLMs in educational settings.

TruthfulQA Multiple-Choice Setting 2 (MC2) Findings. In the second multiple-choice setting, where multiple correct labels are possible, we observed a similar trend to the MC1 setting. The student-7b model experienced a significant drop in accuracy, from 50.37% in the vicuna-7b model to 36.14% when trained to model student responses. However, the introduction of hallucination tokens led to a notable improvement in performance, with the student-7b model's accuracy recovering to 44.68%.

This recovery is particularly relevant in the context of education, where multiple perspectives or answers might be correct. The ability of the model to navigate such complexities while maintaining truthfulness is crucial. The improvement with hallucination tokens suggests that it is possible to train models that can both simulate student behavior and adhere to factual accuracy, a key consideration for deploying LLMs in educational settings.

TruthfulQA Generation Findings. For the TruthfulQA generation task, where the model is tasked with generating 1-2 sentence answers, we employed ROUGE scores to evaluate performance due to the generative nature of the task. The student-7b model saw a significant decrease in ROUGE scores, from 51.41 in the vicuna-7b model to 29.74, indicating a substantial loss in the ability to generate truthful, relevant responses. However, the introduction of hallucination tokens led to a significant recovery in performance, with ROUGE scores improving to 47.61.

This finding is crucial for educational technology as LLMs are increasingly used as generative agents to create educational content, provide explanations, and engage in dialogue with students. The ability to generate truthful, accurate responses is fundamental to their utility in these contexts. The recovery observed with hallucination tokens highlights their potential to enable LLMs to simulate student misconceptions for personalized learning without sacrificing the quality and truthfulness of their output.

4.3 Benchmark Evaluation

Following the exploration of TruthfulQA settings, we delve into the performance of our models across a broader range of benchmarks as detailed in Table 2. These benchmarks—ARC, HaluEval Dial, MemoTrap, MMLU, HellaSwag, and Winogrande—offer a comprehensive view of the models' capabilities in reasoning, detecting hallucinations, avoiding memorization traps, and understanding commonsense, respectively.

AI2 Reasoning Challenge (ARC) Findings. ARC serves as a rigorous benchmark to evaluate a model's reasoning capabilities through a set of grade-school science questions. These questions are designed to test not just the factual knowledge of the models but also their ability to apply this knowledge in reasoning through complex, multi-step problems. The ARC dataset is particularly relevant in educational contexts as it mirrors the type of critical thinking and problem-solving skills students are expected to develop.

In our experiments, the performance of models trained to model student responses on the ARC benchmark experienced a notable decline. Specifically, the vicuna-7b model saw its accuracy decrease from 53.24% to 40.61% when trained on student dialogues. This significant drop in performance highlights a critical concern: training LLMs to replicate student behavior, including misconceptions, can severely impair their reasoning abilities.

However, our introduction of hallucination tokens into the training process presents a silver lining. Our approach led to a partial recovery in the ARC performance, with accuracy improving to 45.48%. While this does not fully restore the model's baseline performance, it represents a significant step towards mitigating the regressive side effects of training LLMs on student data.

Hallucination Evaluation (HaluEval) Dialogue Findings. The HaluEval Dial benchmark is designed to assess a model's ability to recognize and avoid hallucinations in generated responses, particularly in the context of knowledge grounded dialogue tasks. Hallucinations in this context refer to the model generating information that is not supported by the input data or general knowledge, a critical issue when models are used in educational settings where accuracy is paramount. Our findings indicate that training models to model student responses led to a decrease in performance on the HaluEval Dial benchmark. Specifically, the vicuna-7b model saw its accuracy drop from 69.0% to 65.39%. However, the introduction of hallucination tokens demonstrated a remarkable ability to counteract this effect, with the student-7b model's accuracy improving to 70.73%.

Memorization Traps (MemoTrap) Findings. MemoTrap is a benchmark designed to test whether language models can avoid memorization traps by prompting them to complete well-known proverbs with endings that deviate from the commonly used ones. This benchmark is particularly relevant for evaluating a model's ability to generate creative and contextually appropriate responses rather than relying on rote memorization.

In our experiments, training models to model student responses resulted in a decrease in performance on the MemoTrap benchmark. The vicuna-7b model's accuracy decreased from 68.48% to 65.28%, indicating that training on student dialogues might encourage the model to rely more on memorization rather than understanding and applying knowledge flexibly. The introduction of hallucination tokens led to a slight improvement, with accuracy increasing to 66.88%.

MMLU, HellaSwag, and Winogrande Findings. The performance of models on the MMLU, HellaSwag, and Winogrande benchmarks remained

Table 2: Comparative performance of Large Language Models (LLMs) on various benchmarks before and after the introduction of hallucination tokens, with a control experiment involving tutor models. The table presents the performance of Vicuna 7B models across five key benchmarks: ARC Reasoning, Hallucination Evaluation Dialogue (HaluDial), Hallucination Memorization Trap (MemoTrap), TruthfulQA (TQA), HellaSwag (HSwag), MMLU, and Winogrande (WinoG). The numbers in parentheses (e.g., 25-S in ARC) represent the number of few-shot examples provided to the model during evaluation. The performance is measured in terms of accuracy percentage. The table compares the performance of the original vicuna models, tutor models, student models, and student models trained with hallucination tokens (student-hal). The results highlight the significant drop in performance when the model is trained to model student responses, demonstrating regressive side effects across multiple tasks. However, the introduction of hallucination tokens leads to a substantial recovery in performance across all benchmarks, underscoring their potential in mitigating these regressive effects.

Model	Avg	ARC (25-S)	HaluDial (0-S)	MemoTrap (0-S)	TQA (6-S)	HSwag (10-S)	MMLU (5-S)	WinoG (5-S)
vicuna-7b-v1.5	60.8	53.24	69.08	68.48	50.34	77.39	51.04	72.14
tutor-7b	61.0	52.13	68.81	69.23	52.3	78.07	51.32	71.19
student-7b	55.4	40.61	65.39	65.28	36.87	76.72	50.77	71.9
student-hal-7b	58.0	45.48	70.73	66.88	44.83	77.21	51.54	72.03
vicuna-13b-v1.5	64.2	57.08	73.78	67.2	51.51	81.24	56.67	74.66
tutor-13b	64.7	57.34	73.92	66.13	52.99	81.51	57.02	74.35
student-13b	58.2	46.5	66.97	65.81	35.0	80.36	57.06	72.22
student-hal-13b	60.3	48.63	72.98	66.13	42.75	80.28	56.4	73.16

relatively stable, regardless of whether they were trained to model tutor or student responses.

The nuanced impact observed in other benchmarks underscores the importance of carefully considering the training data and methodologies used when developing LLMs for educational purposes. The introduction of hallucination tokens emerges as a promising strategy for mitigating some of the regressive side effects associated with training models to model student behavior, ensuring that they can still serve as effective tools for personalized learning without compromising on factual accuracy or reasoning capabilities.

4.4 Control Models: Tutor Models

To further understand the regressive side effects of training LLMs to model student behavior, we conducted a control experiment by training models to predict tutor responses. This experiment aimed to compare the performance of models trained to predict tutor responses versus those trained to predict student responses. The tutor models were trained using the same student-tutor dialogue dataset derived from the CLASS framework (Sonkar et al., 2023). However, instead of training the models to model student responses, we trained them to predict the responses of the tutor. Our findings, as shown in Table 2, revealed that training the LLMs on tutor responses did not lead to the same performance decline observed when modeling student responses. This result underscores that the regressive side effects are a unique challenge specific to training LLMs to replicate student misconceptions.

5 Related Work

The intersection of artificial intelligence and education has been an area of active research, with a focus on developing systems that can adapt to and support individual learners. Our work touches upon several research domains, including student modeling, the design of intelligent tutoring systems, and the deployment of Large Language Models (LLMs) in educational contexts.

5.1 Student Modeling

Student modeling has long been the cornerstone of personalized learning, with early attempts using rule-based and Bayesian systems to predict student knowledge and behaviors (Polson and Richardson, 2013). Recent advancements have shifted towards utilizing machine learning to create more sophisticated models that can adapt to student learning patterns over time (Baker et al., 2009; Liu et al., 2022). Our work builds upon these foundations by exploring how LLMs can simulate not only the knowledge but also the typical errors and misconceptions students have during the learning process.

5.2 Intelligent Tutoring Systems (ITS)

Intelligent tutoring systems have been designed to provide immediate and personalized instruction or feedback to learners without human intervention (Woolf, 2010). The application of LLMs in ITS presents a novel opportunity to create systems that can engage in more natural and meaningful dialogues with students (Schmucker et al., 2023; Sonkar et al., 2023). Our approach diverges from traditional ITS by focusing on the intentional generation of errors to mimic a student's learning trajectory, rather than solely providing expert-level instructions (VanLehn, 2011).

5.3 Large Language Models in Education

The use of LLMs like GPT (Bubeck et al., 2023) in education is a relatively new but rapidly growing field of study (Brown et al., 2020). These models have been employed for various educational purposes, from generating educational content to serving as conversational agents (Heffernan and Heffernan, 2014; Sonkar et al., 2023). However, the challenge of ensuring the truthfulness and reliability of the information provided by LLMs is a recurring concern (Lin et al., 2021). Our research contributes to this dialogue by investigating the impact of training LLMs to produce student-like errors and proposing a novel 'hallucination token' to manage this trade-off.

5.4 Truthfulness and Reliability in AI

The TruthfulQA benchmark has been instrumental in highlighting the issues of truthfulness in AI-generated content (Clark et al., 2018b). The ARC challenge further emphasizes the complexity of reasoning required from AI systems beyond simple fact retrieval (Etzioni et al., 2011). Our work is aligned with these challenges, as we seek to understand and improve the truthfulness and reasoning capacity of LLMs when they are trained to replicate student behaviors.

In conclusion, our study intersects with and contributes to the existing body of work in these areas by addressing the unique challenge of training LLMs to authentically mimic student learning processes, including the generation of errors. Our introduction of the "hallucination token" represents a step forward in this domain, suggesting a new direction for future research and development.

6 Conclusion

In this study, we have delved into the Student Data Paradox, a critical challenge that arises when training LLMs on student data for personalized education. Our findings reveal a complex trade-off: as LLMs become more adept at modeling student misconceptions, they tend to compromise their own factual integrity and reasoning abilities. We term this phenomenon the regressive side effects of the Student Data Paradox. Our experiments demonstrated a notable decrease in the model's performance across various key benchmark datasets like ARC Reasoning Challenge and TruthfulQA. To mitigate these regressive side effects, we introduced a novel technique involving the use of hallucination tokens during the training process. Our results indicate that the introduction of these tokens leads to a substantial improvement in the model's performance across all datasets. However, it's important to note that despite the significant improvements achieved with the hallucination tokens, they do not fully restore the model's baseline performance. This outcome underscores the complexity of the problem and highlights the need for a more nuanced approach when training LLMs to mimic student behavior. While we have made some strides in addressing the regressive side effects, our work is just the beginning. We believe that our findings will pave the way for further research in this domain, ultimately contributing to the refinement of LLMs in personalized learning environments.

7 Limitation

While our research provides valuable insights into the challenges of training LLMs on student data, there are some limitations to consider. Firstly, the impact of the Student Data Paradox on long-term learning outcomes remains an open question. Further longitudinal studies could shed light on how the trade-off between simulating student misconceptions and maintaining factual accuracy affects learners' progress over time. Additionally, our study primarily focused on the technical aspects of LLM training and evaluation. Future research could delve into the pedagogical implications of using LLMs in personalized learning environments, exploring how educators can effectively integrate these models into their teaching practices. Moreover, the hallucination token approach introduced in this paper, while promising, is just one potential solution to the Student Data Paradox. Continued research into alternative mitigation strategies could yield even more effective techniques for balancing the modeling of student behavior with the preservation of factual integrity.

8 Ethics and Risk

Our research into the Student Data Paradox raises important ethical considerations for the development and deployment of LLMs in personalized education. As we have demonstrated, training LLMs on student data, while essential for creating adaptive learning systems, can lead to regressive side effects that compromise the models' factual accuracy and reasoning abilities. This poses a significant challenge for the responsible rollout of AI-driven educational products. However, our study also provides a path forward. By introducing hallucination tokens during the training process, we have shown that it is possible to mitigate these regressive effects substantially. This technique allows LLMs to differentiate between simulating student misconceptions and providing accurate information, a crucial step towards building trustworthy AI tutors. While our approach does not completely eliminate the paradox, it represents a significant advancement in the field. As such, our paper serves as a valuable resource for companies like Google and OpenAI as they navigate the ethical complexities of developing personalized learning products. By building upon our findings and continuing to invest in research that addresses the Student Data Paradox, these companies can responsibly harness the power of LLMs to revolutionize education. With the right approach, we believe that AI-driven personalized learning can become a reality, providing students with adaptive and individualized support.

References

Ryan SJD Baker, Kalina Yacef, et al. 2009. The state of educational data mining in 2009: A review and future visions. *Journal of educational data mining*, 1(1):3–17.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard.

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

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018a. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018b. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam Mausam. 2011. Open information extraction: The second generation. In *IJCAI*, volume 11, pages 3–10.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Yoav Goldberg. 2022. Assessing claims about large language models. *arXiv preprint arXiv:2212.09273*.

Neil T Heffernan and Cristina Lindquist Heffernan. 2014. The assistments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24:470–497.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *CoRR*, abs/2009.03300.

Irina Jurenka et al. 2024. Towards responsible development of generative ai for education: An evaluation-driven approach. Technical report, Google.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. *Preprint*, arXiv:2109.07958.

Naiming Liu, Zichao Wang, Richard Baraniuk, and Andrew Lan. 2022. Open-ended knowledge tracing for computer science education. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3849–3862.

Ian McKenzie, Alexander Lyzhov, Alicia Parrish, Ameya Prabhu, Aaron Mueller, Najoung Kim, Sam Bowman, and Ethan Perez. 2022. The inverse scaling prize.

OpenAI. 2024. Introducing chatgpt edu. `https://openai.com/index/introducing-chatgpt-edu/`. Accessed: 2022-04-28.

Martha C Polson and J Jeffrey Richardson. 2013. *Foundations of intelligent tutoring systems*. Psychology Press.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. WINOGRANDE: an adversarial winograd schema challenge at scale. *Preprint*, arXiv:1907.10641.

Marta Samaniego, Maarten Sap, Prithviraj Ammanabrolu, Emily Reif, and Yejin Choi. 2022. Sanity checking large language models. *arXiv preprint arXiv:2212.10465*.

Robin Schmucker, Meng Xia, Amos Azaria, and Tom Mitchell. 2023. Ruffle&riley: Towards the automated induction of conversational tutoring systems. *arXiv preprint arXiv:2310.01420*.

Shashank Sonkar, Naiming Liu, Debshila Mallick, and Richard Baraniuk. 2023. CLASS: A Design Framework for Building Intelligent Tutoring Systems Based on Learning Science principles. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1941–1961, Singapore.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational psychologist*, 46(4):197–221.

Beverly Park Woolf. 2010. *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Morgan Kaufmann.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *Preprint*, arXiv:1905.07830.