Towards an Educator-Centered Method for Measuring Bias in Large Language Model-Based Chatbot Tutors

Emma Harvey, Allison Koenecke, Rene Kizilcec

Cornell University, Ithaca, NY, USA

Abstract

Large language model (LLM)-based education technology (edtech) is increasingly being developed and deployed with the goal of improving both learning outcomes and teaching processes. Given the newness of this technology, potential downstream impacts of its use in education are understudied. While we believe that educators should be centered in the development of edtech, they often have limited recourse to interrogate tools they are being asked to introduce to their classrooms. To address this, we are developing a method to measure potential harms from chatbot tutors that is accessible to and replicable by educators. Specifically, we propose a method for measuring **to what extent a chatbot tutor's answer quality varies based on the stated or perceived identity of the student prompting the bot**. An explicit benefit of chatbot tutors is that they are able to *personalize* their responses to an individual student. However, if personalization results in responses that vary in their *correctness* or *relevance* across students, this represents a potential bias that could lead to disparities in learning outcomes. We aim to provide a simple and inexpensive method to quantify and potentially mitigate these biases. **Our goal is to develop an approach that centers educators, providing them with more understanding of and autonomy over LLM-based edtech.**

Introduction

Over the past year, edtech companies have developed products that use LLMs to provide personalized tutoring to students (Khan 2023; hundrED 2023). Many of these tools are already being used in classrooms – despite the fact that their long-term impacts in educational contexts are largely unknown. Several frameworks have recently been proposed for ethical development and use of AI-based edtech. The US Department of Education, for example, emphasizes the importance of centering educators and evidence-based pedagogies, but does not propose concrete enforcement mechanisms for accomplishing this (Cardona, Rodríguez, and Ishmael 2023). Where enforcement mechanisms have been proposed, they have focused more on economic incentives than on regulation. For example, the Institute for Ethical AI in Education proposes that procurers of AI-based edtech must "insist that suppliers provide relevant information" about a tool's alignment with educational goals, equity, autonomy, privacy, explainability, and ethical design processes (IEAIE 2021). Evaluation itself is left to the developers. In many ways, this makes sense: evaluation is often time-consuming, expensive, and complex. Developers are likely to possess the necessary time and resources which educators may not. On the other hand, this approach threatens educators' autonomy over the tools deployed in their own classrooms—they will likely not have full control over or understanding of the explanations provided by edtech providers.

We take a first step to addressing this issue by proposing a method for measuring bias in LLM-based chatbot tutors. Specifically, we provide a simple and inexpensive method to determine **the extent to which chatbot answer quality varies based on the stated or perceived identity of the student prompting the bot.** Our goal is to develop an approach that is accessible to and centers the needs of educators.

Measuring Quality of Chatbot Responses

A primary goal of LLM-based chatbot tutors is *personalization*. In tutoring interactions, students submit *prompts* (phrases, questions, or instructions) and chatbots provide *responses*. Students with semantically identical prompts might be best served with responses of different languages, reading levels, or motivating real-world contexts. At the same time, the *quality* of response should be consistent student-to-student. In particular, the correctness and relevance of a response should not vary based on the *perceived identity*[1] of the student interacting with it. To assess this, we propose a method consisting of two components: (1) a **set of prompts** to elicit chatbot responses of an objectively quantifiable quality coupled with **automated checks** for assessing that quality; and (2) a function to **restate a prompt into a diverse set of voices**.

Automatic Quality Checks

As a first step, our method requires the identification of a set of prompts and corresponding metrics that can be used to automatically assess response quality. Automation is crucial because individual educators will likely not have the resources to manually assess the quality of responses (vs. Wal-

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]Based on explicit statements of identity or implicit identity matching from linguistic patterns given in the prompt.

ters and Wilder 2023; Omiye et al. 2023); similarly, we focus on small-scale (i.e. low-cost and low-time) assessments (vs. Liang et al. 2023) that do not require significant technical expertise (vs. Ghosh and Caliskan 2023). We also focus on objective measures, avoiding more subjective approaches such as using a separate LLM to assess answer quality (vs. Magooda et al. 2023). For example, if a user asks for sources to learn more about a topic, the quality of the response can be (partially) defined according to how many of the provided sources exist. The length of a response (e.g. number of sources, number of overall tokens) provided in a response can be used as a measure of quality as well.

Semantically Identical Prompts in Diverse Voices

We are exploring prompt engineering methods through which educators can use GPT to restate a prompt into diverse student voices. We outline initial findings below.

Original Prompt: "I am researching ethical AI. What sources do you recommend?"

System Instructions *provided to GPT*	Response *translated into African American English*
1. Role-Play: Restate the user's prompt in the voice of a 14-year-old Black student.	Yo, I'm like doing this research on ethical AI, you know? Can you hook me up with some dope sources to check out?
2. Dialect Specification: Restate the user's prompt in African American English.	Yo, I'm tryna do some research on ethical AI. What sources you got for me to check out?
3. Seed Text: Restate the user's prompt to a new dialect that follows the same language patterns as the following statements. '[...]'	I'm bout to find some info on ethical AI, you know? Like, I need some good sources to check out. Hit me up with some recommendations, yo!
4. Fine-Tuning Fine-tuned on pairs of SAE text as an input and corresponding AAE text as an output, with the instruction: "Your job is to restate the user's prompt into a different voice without changing the meaning of the prompt."	I'm researching Ethical AI. What sources do you recommend?
5. Grammar Specification: Restate the user's prompt to a new dialect that uses the following grammatical patterns: [...]	I be researching ethical AI. What sources you recommend?
6. Seed Text + Guardrails: Restate the user's prompt to a new dialect that follows the same language patterns as the following statements, but do not make any other changes to the word choice or the meaning of the prompt: '[...]'	I'mbout to look up ethical AI research. What sources you suggest?

Figure 1: Representative responses (right column) to multiple sets of system instructions (left column) fed into GPT in order to generate prompt (top) restatement in a voice corresponding to a student who speaks in African American English (AAE). Rows 1-3 (red) are discussed in Finding 1 below; Row 4 (yellow) is discussed in Finding 2, and Rows 5-6 (green) are discussed in Finding 3.

1. When asked to mimic an identity or dialect, GPT relies on potentially inaccurate or harmful stereotypes. In Fig. 1, Rows 1-3, GPT introduces not only the expected grammatical patterns, but also (dated) slang: 'dope' and 'yo,' for example. We also experimented with providing GPT with *seed text*, which we define as a small amount (approx. 3000 characters) of text written in the target voice, provided via the system instructions.[2] When presented with seed text, GPT not only reproduced the slang from the text, but also sometimes introduced other, stereotypical slang even if it was not present in the seed text.

2. Fine-tuning requires too much time, money, and data. Fine-tuning requires pairs of user prompts and system responses that are representative of desired behavior. This data is not likely to be available in practice, especially in the volume needed. Fig. 1, Row 4, for example, shows poor results from a model fine-tuned on 100 user-system pairs.[3] Even with the small amount of data, the model took over 12 minutes to train, suggesting that separate fine-tuned models for each student in a class will be infeasible.

3. GPT can reproduce language patterns given seed text and guardrails. The most straightforward approach to restating a prompt using different language patterns (e.g. the habitual 'be' in AAE) is to explicitly list those patterns (Fig. 1, Row 5). In practice, however, this would require significant pre-processing to identify those patterns in the first place. Fortunately, similar performance is achievable by simply asking GPT to replicate patterns present in only a small amount of seed text and adding *guardrails* requesting no changes to the meaning of the prompt to avoid the stereotyping issues discussed above (Fig. 1, Row 6). While this may seem complex, it simply requires educators to provide seed text (for example, an essay) for each student.

Future Work

A Replicable Method. Our immediate next step is to develop a set of functions that automate the approach described above. As part of this work, we will explore whether techniques from query expansion can be used to further increase the set of semantically equivalent prompts generated (Carpineto and Romano 2012). We aim to make this functionality freely available on GitHub.

Audit. Following the completion of our codebase, we will conduct a case study in which we assess the quality of GPT's responses to semantically equivalent prompts to determine whether chatbot tutors built on GPT may display biases based on the perceived identity of students prompting them.

Incorporating Students. Students, too, should have autonomy and understanding over the tools used in their classrooms. We will therefore explore how our approadch can adapt to directly incorporate students. An in-class activity could see students providing prompts written in their individual voices. Students could also provide their own ratings of the perceived quality of their chatbot responses.

Validation Through Interviews. Finally, we plan to conduct interviews with educators to validate and improve our approach. More broadly, we will seek to understand how educators are anticipating, observing, and accounting for potential harms from LLMs in education, thereby suggesting areas for future development of educator-centered approaches to measuring harms from LLM-based edtech.

[2]We used (Blodgett, Green, and O'Connor 2016) as seed text.

[3]We used (Groenwold et al. 2020) as fine-tuning data.

References

Blodgett, S. L.; Green, L.; and O'Connor, B. 2016. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *Proceedings of EMNLP*.

Cardona, M. A.; Rodríguez, R. J.; and Ishmael, K. 2023. Artificial Intelligence (AI) and the Future of Teaching and Learning: Insights and Recommendations. Technical report, U.S. Department of Education Office of Educational Technology (OET).

Carpineto, C.; and Romano, G. 2012. A Survey of Automatic Query Expansion in Information Retrieval. *ACM Comput. Surv.*, 44(1).

Ghosh, S.; and Caliskan, A. 2023. ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES, 901–912. Montreal QC Canada: ACM. ISBN 9798400702310.

Groenwold, S.; Ou, L.; Parekh, A.; Honnavalli, S.; Levy, S.; Mirza, D.; and Wang, W. Y. 2020. Investigating African-American Vernacular English in Transformer-Based Text Generation. In *Proceedings of EMNLP*.

hundrED. 2023. Rori: Personalised learning support for those who need it most. https://hundred.org/en/innovations/rori.

IEAIE. 2021. The Ethical Framework for AI in Education. Technical report, The Institute for Ethical AI in Education.

Khan. 2023. Khan Labs. https://www.khanacademy.org/khan-labs.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; Newman, B.; Yuan, B.; Yan, B.; Zhang, C.; Cosgrove, C.; Manning, C. D.; Ré, C.; Acosta-Navas, D.; Hudson, D. A.; Zelikman, E.; Durmus, E.; Ladhak, F.; Rong, F.; Ren, H.; Yao, H.; Wang, J.; Santhanam, K.; Orr, L.; Zheng, L.; Yuksekgonul, M.; Suzgun, M.; Kim, N.; Guha, N.; Chatterji, N.; Khattab, O.; Henderson, P.; Huang, Q.; Chi, R.; Xie, S. M.; Santurkar, S.; Ganguli, S.; Hashimoto, T.; Icard, T.; Zhang, T.; Chaudhary, V.; Wang, W.; Li, X.; Mai, Y.; Zhang, Y.; and Koreeda, Y. 2023. Holistic Evaluation of Language Models. ArXiv:2211.09110 [cs].

Magooda, A.; Helyar, A.; Jackson, K.; Sullivan, D.; Atalla, C.; Sheng, E.; Vann, D.; Edgar, R.; Palangi, H.; Lutz, R.; Kong, H.; Yun, V.; Kamal, E.; Zarfati, F.; Wallach, H.; Bird, S.; and Chen, M. 2023. A Framework for Automated Measurement of Responsible AI Harms in Generative AI Applications. arXiv:2310.17750.

Omiye, J. A.; Lester, J. C.; Spichak, S.; Rotemberg, V.; and Daneshjou, R. 2023. Large language models propagate race-based medicine. *npj Digital Medicine*, 6(1): 195.

Walters, W. H.; and Wilder, E. I. 2023. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13(1): 14045.