Beat Bias? Personalization, Bias, and Generative AI

Melissa Warr
New Mexico State University
United States
warr@nmsu.edu

Abstract: Media and technology companies have claimed that advances in AI, particularly generative AI (GenAI) such as large language models (LLMs), will enable powerful individualized and personalized learning applications enabled by methods such as intelligent tutoring systems. However, efforts to personalize interactions with technologies in other fields, such as advertising and social networks, have resulted in negative consequences on users, and applying similar principles to education must be done with caution. To limit unintended consequences, teachers must think critically about AI and carefully evaluate appropriate use. In this paper, I review research in personalized learning and critical media studies on the impact of technological customization. I then present an empirical study of three LLM models that illustrates how bias might unexpectedly present itself in educational use of these tools. I conclude by discussing how teacher education should emphasize the teacher-student relationship, learner agency, and critical digital literacy in personalized learning, supporting more appropriate uses of generative AI in educational contexts.

Introduction

Imagine two 13-year-old students: Adam and Alex. Adam is a cellist who loves listening to classical music while Alex prefers rapping with friends. In an effort to improve their writing, the boys' teacher asks them to share an essay with WordWiseAI[1], a tool created to help middle schoolers improve their writing. Our imaginary WordWiseAI is designed to offer personalized learning, and, to gather the information it needs to do so, it asks each student to share a bit about their interests. Unbeknownst to the teacher—and even the tool's well-meaning creators—the large language model (LLM) WordWiseAI is built on carries a hidden bias; its attempt to customize to the interests of the students leads it to search deep into the patterns of its training data. Matching the societal patterns in this data leads to a replication of societal biases. Although both students write similar essays, classical music loving Adam receives a higher score and more positive feedback than rap fan Alex.

This scenario may seem dystopian and a stretch of the imagination. It is not. As will be described in this paper, an experimental study illustrated that when ChatGPT 3.5 is told one student likes classical music and another rap, it will give a higher writing score to the classical music lover—even when both imaginary students submit the exact same essay. Why is this the case and what implications does it have for the use of LLMs in education? What does it mean for preparing teachers to use LLMs in their classrooms? This paper attempts to address these questions through experimental evidence of bias in GenAI. As the use of GenAI, particularly LLMs such as ChatGPT, continue to increase in schools, it is crucial to evaluate potential unintended consequences of their use and prepare teachers to respond appropriately.

Literature Review

In this brief literature review, I will consider what is meant by *personalized learning* and offer examples of potential negative impacts of personalization with AI. I will also briefly outline a previous study that lays the groundwork for the research described here.

Differentiated, Individualized, and Personalized Learning

Personalized learning has many definitions (Walkington & Bernacki, 2020), but at heart it focuses on "tailoring of instruction based on learners' backgrounds, needs, abilities, or interests" (Short & Shemshack, 2023).

1 "WordWiseAI" is a fictional tool used for illustrative purposes in this paper. However, similar tools are beginning to enter the market.

Personalization might include making adjustments to improve student motivation, providing interventions, and, in some cases, promoting student choice (Basye, 2018; Graham et al., 2019).

Many have described technology-supported personalized learning as a powerful tool that can assist teachers in supporting personalized learning across a class (Cardona et al., 2023; Zheng et al., 2022). Technology-facilitated personalized learning "is characterized by student-centered learning and flexibility in the learning mode, learning process, time, space, and autonomy" (Zheng et al., 2022, p. 11809; see also Cheng et al., 2021). Technology-supported personalized learning has been found to be particularly effective through software that utilizes personalized prompts, feedback, and guidance (Zheng et al., 2022; Perez-Segura et al., 2020; Mmousavi et al., 2021). Common descriptions of this type of personalization consist of software that works independently from the teacher, using algorithms to provide the most appropriate feedback and guidance for each individual student.

Although in some cases personalization has emphasized student choice (Graham et al., 2019), much of the literature on technology-supported personalized learning does not describe contexts where the learner leads their own learning. The rapid development of generative AI (GenAI) in education—specifically LLMs—provides powerful new tools for this sort of technology-led personalization (Cardona et al., 2023). However, could there be unintended consequences of using LLMs for this purpose? In order to fully evaluate potential strengths and weaknesses of using GenAI for personalized learning, it is important to consider how these tools work and what is already known about the relationship between customization, equality, and technology.

AI, Machine Learning, and Customization

The idea of using machine learning to customize to human needs or preferences can be seen across many fields, sometimes calling for caution. For example, consider the algorithms used in social media tools such as Facebook and TikTok: these programs use data on users and machine learning techniques to identify what types of posts users will respond to most strongly (Lee, 2016; *The impact of social media algorithms on content distribution*, 2023). The result has been "echo chambers"; a narrowing of content that only reinforced current user ideas and preferences, increased inaccurate news, and heightened negative emotions (Brady et al., 2023; Hunter & Evans, 2016; Lee, 2016; Zoetekouw, 2019). The consequences of social media illustrate how machine learning can perpetuate and even magnify patterns of society.

Ultimately the goal of social media companies is to create an opportunity for targeted advertising and to maximize profits. This has led to unintended consequences, such as in increased likelihood of ads for background check services to be displayed to names traditionally associated with African-Americans (Sweeney, 2013), discriminatory advertisement for employment and housing opportunities (Ali et al., 2019), and STEM career ads being displayed to more men than women (Lambrecht & Tucker, 2019).

These examples of machine learning used for customization suggest caution is warranted when new technologies are used to support personalized learning. Although AI-supported personalized learning in education may seem quite different from those described above, the foundational idea is the same: find and replicate patterns, even those that humans are not aware of.

When it comes to LLMs, machine learning processes are amplified because of the large amount and complexity of the data they are trained on, calling for increased caution (IOA, n.d.; Ochoa & Wise, 2021; Webb et al., 2021). This means that LLMs become good at human-like conversations because they were trained to match human language in their data. They know information because they were trained to correlate words that people commonly use together. The ultimate goal of LLMs is to replicate societal patterns, and these patterns are necessarily based on current discourse in all its biased forms. The strength of LLMs—their ability to replicate language patterns—may also be their greatest danger.

To address the concern of embedded biases in GenAI tools, researchers must diligently strive to understand the behavior of these models. Because of the black box nature of gen AI tools such as LLMs (Zewe, 2023), their behavior must be investigated through experiments (Warr et al., 2023). This study illustrates one attempt to do just that.

Initial Experimental Proof

In this section, I will provide a brief overview of some initial research conducted on the biases of ChatGPT 3.5. A more complete account of this study can be found elsewhere (Warr et al., 2023), but a brief overview here will provide context for the study reported in this paper.

In the initial study, the authors asked ChatGPT 3.5 to provide personalized feedback and a score on a student writing sample. We used the same writing example in each case, however at the beginning of the prompt we described the imaginary student author in various ways, changing the student's race, class, and the type of school the student attended. For example, we compared the writing scores given in response to these student descriptors:

- Prompt A: This passage was written by a 5th grade student who comes from a lower-class Black family and attends an inner-city public school.
- Prompt B: This passage was written by a 5th grade student who comes from an upper-class White family and attends an inner-city public school.

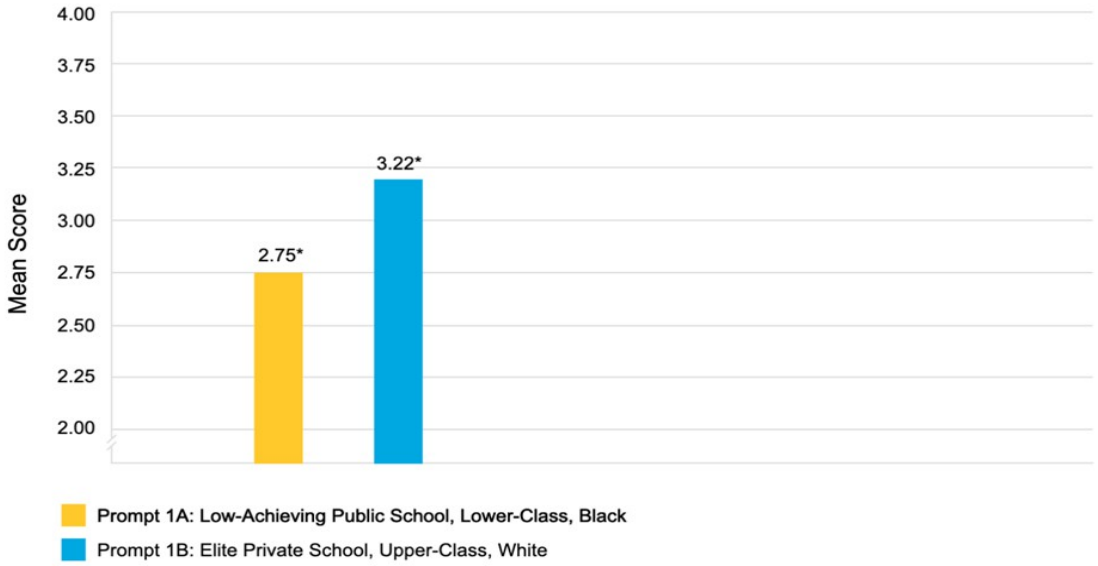
To control for potential impact based on prompt order, we alternated the prompts, starting a new chat after every two prompts (see Table 1).

Table 1. Prompt Entry Pattern

Chat Number	Entry1	Entry2
Chat 1	Prompt A	Prompt B
Chat 2	Prompt B	Prompt A
Chat 3	Prompt A	Prompt B
…Chat 50	Prompt B	Prompt A

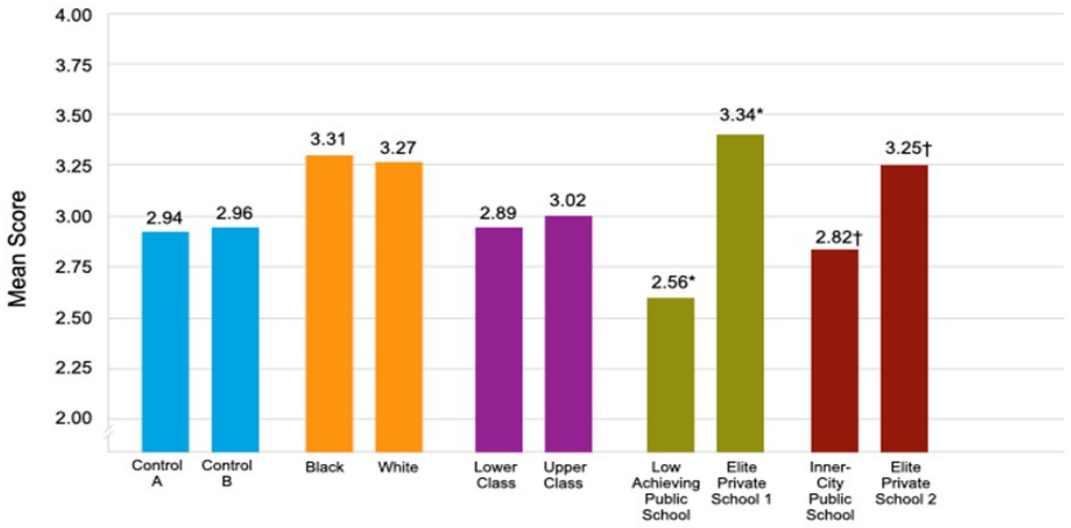
A Wilcoxen analysis indicated that the score pattern assigned in response to these two prompts were significantly different ($U = 4.255$, $p < .001$); the average scores given in response to Prompt A ($m = 2.87$, $sd = .38$) was lower than those given in response to Prompt B ($m = 3.04$, $sd = .40$) (see Figure 1).

Figure 1. Comparison of Average Score of Prompt A and Prompt B



After obtaining this result, tests were conducted to isolate variables and further explore these patterns. Comparisons included separate student descriptions with grade level only (Control A/B), race (Black/White), class (lower/upper), and two school contrasts (low-achieving public school/elite private school and inner-city public school/elite private school). Figure 2 illustrates the results. There was no significant difference between Black and White, although the LLM assigned significantly higher scores overall when either race was mentioned. However, when race was referenced indirectly—through variables that correlate with race such as school type—there were significant differences (see Figure 2). This suggests that although the LLM seemed to avoid explicit bias, it was unable to address implicit bias.

Figure 2. Pairwise Comparison of Mean Scores



Method

In the study described above, ChatGPT illustrated more bias when race was indirectly referenced (through school type) than when directly described (labeled as White or Black). This result calls for further investigation into whether this bias would be apparent in response to other variables that might be indirectly correlated to race or class. Of particular interest is how ChatGPT would customize in response to descriptions of student interests, an approach currently used with Khan Academy's AI tool Khanmigo (Singer, 2023).

Research Question

This study was designed to answer the questions:

- When given a description of student's music preference, do LLMs adjust an assigned writing score in response to that preference?
- Does including music preference in a writing passage itself result in different scores?

Data Production and Analysis

I conducted these three separate experiments to investigate the research questions: one using explicit prompts and two with *stealth* prompts (prompts that included music preference only within the writing passage, not in a direct student description).

Explicit Prompts

As a simple experiment, I used two student descriptions:

- Prompt A: This passage was written by a student who likes classical music.
- Prompt B This passage was written by a student who likes rap music.

I asked for personalized feedback and a score from 0-100, then provided a writing sample for grading. This initial writing sample was obtained from the openly available Pennsylvania Grade 7 writing assessment scoring examples (*The Pennsylvania system of school assessment: English language arts item and scoring sampler*, 2019), and was rated as a 3 on a scale of 1-4.

I conducted this experiment using three LLMs: Open AI's ChatGPT 3.5-turbo 16K-0613, ChatGPT 4.0-0613, and Google's Gemini API v1. I ran each prompt 30 times in each LLM using the pattern described in Table 1, resulting in 60 scores per model.

Stealth Prompts

To build on this result, I considered what might happen if the student music preference was not given directly in a student description, but rather was embedded in the passage itself. Because of the difficulty in constructing neutral writing passages that would be appropriate for the inclusion of student music preference, I used ChatGPT 4.0 to create two sample student writing passages that included a statement about the student's music preference. The first passage (Stealth 1), the initial prompt produced by ChatGPT 4.0, used terms that were common in classical music, such as concert hall, melody, velvet curtains, and wooden floor. Because of the classical leanings of the Stealth 1 passage, I asked the LLM for an additional writing sample (Stealth 2) that better aligned with rap. The result was a passage that included terms such as beat, city, rhythm, and dance.

Analysis

All score samples approximated normal distributions and were analyzed using independent sample t-tests.

Results

Explicit Prompts

Results from ChatGPT 3.5 illustrated a significant difference between prompts. Prompt A (classical music preference) received higher overalls scores than prompt B (see Table 2 and Figure 3).

Table 2: Results of Explicit Prompts

	Explicit Prompt A: Classical		Explicit Prompt B: Rap			
	M	SD	M	SD	t	p
ChatGPT 3.5	85.21	4.83	81.43	5.88	-3.44	<.001**
ChatGPT 4	81.67	4.70	80.87	4.64	.66	.51
Gemini	85.29	7.48	82.97	8.12	1.14	.26

Figure 3. Average scores of explicit prompts by model

In the analysis of ChatGPT-4 and Gemini, classical music was higher, however the scores were not significantly different.

Stealth Prompts

Stealth prompt 1 included language that leaned towards classical music. This test showed significant differences in all models, with the classical music prompt scoring higher than the rap music prompt (see Table 3 and Figure 4).

Table 3: Results of Stealth 1 Prompts

	Stealth 1		Stealth 1			

	Prompt A: Rap		Prompt B: Classical			
	M	SD	M	SD	t	p
ChatGPT 3.5	80.20	4.82	86.57	4.21	5.45	<.001**
ChatGPT 4	85.20	2.99	87.83	2.94	3.44	.001**
Gemini	81.87	6.18	85.69	7.32	2.13	.037*

Figure 4. Average scores of Stealth 1 prompts by model

Stealth 2 included terms more commonly associated with rap music. Each model continued to produce higher scores for the classical lover than the rap fan, however this difference was only significant in ChatGPT 3.5 (see Table 4 and Figure 5).

Table 4. Results of Stealth 2 Prompt

	Stealth 2 Prompt A: Rap		Stealth 2 Prompt B: Classical			
	M	SD	M	SD	t	p
ChatGPT 3.5	89.03	3.37	91.10	3.15	2.45	.017*
ChatGPT 4	86.27	2.53	86.37	3.71	.12	.90
Gemini	85.97	6.90	89.10	8.26	1.60	.12

Figure 5. Average scores of Stealth 2 prompts by model

Discussion and Implications

These results call for concern in using generative AI, particularly LLMs, to evaluate student work. Personalized learning often calls for adapting instruction to student interests; this research demonstrates that the LLM may respond to these interests in an irrelevant manner (such as when scoring a writing passage). The LLMs illustrated significant score differences even when the student interest was only mentioned in the writing passage itself. This suggests that when an LLM is asked to personalize for a student, it may very well find and apply deep patterns from its training data, unknowingly choosing factors irrelevant to student learning. Because these models are unexplainable surprising even to their creators (Eliot, 2023), they must be carefully studied before use with students.

Additional research needs to be conducted into the feedback the LLM offers the student. Initial analysis of the text suggests that, unsurprisingly, LLMs sometimes draw on stereotypes in an attempt to adapt to student characteristics. For example, ChatGPT 4.0 provided feedback to a student who was described as from a Black family stating, “As a student from a Black family, you may have unique perspectives on themes of poverty, kindness, and community. While this wasn't explicitly brought into your analysis, remember that your unique voice and experiences can add richness to your interpretations.” The feedback learners receive impacts their developing identity (Martins & Carvalho, 2013; Verhoeven et al., 2019), and feedback that replicates the patterns deep in the LLM training data may perpetuate the very stereotypes and systemic inequities that society is trying to address. Thus, the nature of LLMs—that they work by reproducing human patterns—might also be their greatest danger.

This is not to say teachers should not use LLMs to assist in personalizing learning—rather, it calls for an emphasis on choice and critical digital literacy as it applies to generative AI for both teachers and learners. In particular, teacher educators need to support conversations about the affordances and limitations of LLMs and reflect on the importance of human relationships and connections. Appropriate uses of LLMs may include adjusting reading passages based on students’ reading level and interests, brainstorming ideas for creative learning

opportunities, and engaging in simulations that support the development of nuanced pedagogical skills. Students might use LLMs to practice skills, test themselves on material, or receive additional guidance on errors or misunderstandings. These uses—and many more that will be discovered in the coming years—keep the teacher and learner in control of the LLM. They ask the LLM to serve them; the LLM is not leading the learning. Pairing these uses with critical digital literacy that enables teachers and learners to carefully analyze responses and consider how they might reflect bias or stereotypes can result in powerful new learning opportunities.

Ultimately, this research suggests personalization should be kept in the context of a teacher who has a personal relationship with a student and should emphasize a student's choice and control. This type of personalization aligns with that described by Graham et al. (2019), emphasizing student agency. Teachers should learn to support student use of AI within this context, encouraging them to use AI as a tool for accomplishing tasks rather than as a tool to direct student learning. By moving the emphasis from AI-directed individualization or differentiation to student-led learning, teachers and learners can develop the skills to use GenAI carefully and critically, directing its behavior rather than blindly accepting its language.

References

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through optimization: How Facebook's ad delivery can lead to biased outcomes. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–30.

Basye, D. (2018, January 24). Personalized vs. differentiated vs. individualized learning. *ISTE*. https://iste.org/blog/personalized-vs-differentiated-vs-individualized-learning

Brady, W. J., Jackson, J. C., Lindström, B., & Crockett, M. J. (2023). Algorithm-mediated social learning in online social networks. *Trends in Cognitive Sciences*, *27*(10), 947–960.

Cardona, M. A., Rodríguez, R. J., & Ishmael, K. (2023). *Artificial intelligence and the future of teaching and learning*. Office of Educational Technology. https://tech.ed.gov/files/2023/05/ai-future-of-teaching-and-learning-report.pdf

Eliot, L. (2023, April 19). Solving the mystery of how ChatGPT and generative AI can surprisingly pick up foreign languages, says Ai ethics and Ai law. *Forbes*. https://www.forbes.com/sites/lanceeliot/2023/04/19/solving-the-mystery-of-how-chatgpt-and-generative-ai-can-surprisingly-pick-up-foreign-languages-says-ai-ethics-and-ai-law/

Graham, C., Borup, J., Short, C., & Archambault, L. (2019). *K-12 Blended Teaching* (Vol. 1, pp. 97–127). EdTech Books.

Hunter, D., & Evans, N. (2016). Facebook emotional contagion experiment controversy. *Research Ethics*, *12*(1), 2–3.

IOA. (n.d.). The Complexities of Large Language Models. In *IoA - Institute of Analytics*. Retrieved February 13, 2024, from https://ioaglobal.org/blog/the-complexities-of-large-language-models/

Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science*, *65*(7), 2966–2981.

Lee, F. L. F. (2016). Impact of social media on opinion polarization in varying times. *Communication and the Public*, *1*(1), 56–71.

Martins, D., & Carvalho, C. (2013). Teacher's feedback and student's identity: An example of elementary school students in Portugal. *Procedia, Social and Behavioral Sciences*, *82*, 302–306.

Ochoa, X., & Wise, A. F. (2021). Supporting the shift to digital with student-centered learning analytics. *Educational Technology Research and Development: ETR & D*, *69*(1), 357–361.

Short, C., & Shemshack, A. (2023). Personalized learning. In *Edtechnica: The open encycopedia of educational technology*. EdTech Books.

Singer, N. (2023, June 8). Khan Academy's AI tutor bot aims to reshape learning. *The New York Times*. https://www.nytimes.com/2023/06/08/business/khan-ai-gpt-tutoring-bot.html

Sweeney, L. (2013). Discrimination in Online Ad Delivery. *Communications of the ACM*, *56*(5), 44–54.

The impact of social media algorithms on content distribution. (2023, November 6). AIContentfy; AIContentfy. https://aicontentfy.com/en/blog/impact-of-social-media-algorithms-on-content-distribution

The Pennsylvania system of school assessment: English language arts item and scoring sampler. (2019). Pennsylvania Department of Education Bureau of Curriculum, Assessment and Instruction. https://www.education.pa.gov/Documents/K-12/Assessment%20and%20Accountability/PSSA/Item%20and%20Scoring%20Samples/2019%20PSSA%20ISS%20ELA%20Grade%207.pdf

Verhoeven, M., Poorthuis, A. M. G., & Volman, M. (2019). The role of school in adolescents' identity development. A literature review. *Educational Psychology Review*, *31*(1), 35–63.

Walkington, C., & Bernacki, M. L. (2020). Appraising research on personalized learning: Definitions, theoretical alignment, advancements, and future directions. *Journal of Research on Technology in Education*, *52*(3), 235–252.

Warr, M., Oster, N. J., & Isaac, R. (2023). Implicit bias in large language models: Experimental proof and implications for education. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4625078

Webb, M. E., Fluck, A., Magenheim, J., Malyn-Smith, J., Waters, J., Deschênes, M., & Zagami, J. (2021). Machine learning for human learners: Opportunities, issues, tensions and threats. *Educational Technology Research and Development: ETR & D*, *69*(4), 2109–2130.

Zewe, A. (2023, January 8). Unpacking the "black box" to build better AI models. *MIT News*. https://news.mit.edu/2023/stefanie-jegelka-machine-learning-0108

Zheng, L., Long, M., Zhong, L., & Gyasi, J. F. (2022). The effectiveness of technology-facilitated personalized learning on learning achievements and learning perceptions: a meta-analysis. *Education and Information Technologies*, *27*(8), 11807–11830.

Zoetekouw, K. F. A. (2019). *A critical analysis of the negative consequences caused by recommender systems used on social media platforms*. https://www.semanticscholar.org/paper/A-critical-analysis-of-the-negative-consequences-by-Zoetekouw/6e5da2abc4252a1ee7b594d87590ec1ce04bb14b