The Ethical Concerns Surrounding GPT in Education

Elham Tajik

1Affiliation not available

January 2, 2024

Abstract

The GPT language model, developed by OpenAI, has gained widespread attention for its impressive ability to engage in conversational interactions with human users, thanks to its extensive training on vast amounts of written data. However, it is crucial to acknowledge that GPTs and their derivatives can exhibit unethical behaviors. Despite this, there has been limited research conducted on the ethical implications of Large Language Models (LLMs) in higher education. To address these concerns, this paper aims to comprehensively examine the harmful behaviors displayed by GPT and its spin-offs. These behaviors encompass social and monolingual biases, issues of reliability and trustworthiness, data privacy and security concerns, toxicity in generated content, challenges in human-computer interaction, and environmental impact. By conducting this study, valuable insights will be provided to educators, administrators, and students, enhancing their awareness of the ethical risks inherent in GPT and its various versions. This will enable them to proactively address and mitigate potential ethical concerns arising from the use of LLMs in educational settings.

Introduction

Language Large Models (LLMs) have attracted significant research attention across various disciplines, extending beyond computer science. For instance, LLMs find applications in medical contexts (1), higher education (2), and other fields, facilitating creative generation such as paraphrasing, summarizing, writing articles, and poetry, as well as supporting decision-making through informed judgments based on data understanding (Zhou et al., 2023).

Despite their wide-ranging applications, LLMs can sometimes exhibit unethical and harmful behaviors, leading to negative biases and generalizations towards specific groups based on factors like gender, race, religion, and social constructs (3). Notable examples include Microsoft’s chatbot Tay, which quickly displayed toxic behavior by spreading racist, sexist, and abusive language, highlighting the risks associated with uncontrolled biases in LLMs (4). Similar behavior was observed in Meta’s chatbot, Galactica’s function (5). Research has also revealed that LLMs like BERT (6), ROBERTA (7), XLNET (8), and GPT-2 (9) exhibit strong stereotypical biases related to race, gender, religion, and profession (10).

In the realm of political statements, Hartmann, et al. found that GPT demonstrated a left-liberal orientation, indicating a pro-environment stance (11). Additionally, Krügel, et al. observed that GPT lacks consistency as a moral advisor, influencing human moral decision-making (12). GPT-3 has been shown to perpetuate enduring stereotypical biases against Muslims (13, 14) and exhibit gender stereotypes during narrative generation (15).

Scholars in computer science, such as Weidinger, et al. (16) and Zhuo et al. (17) have created a taxonomy of unethical behaviors within LLMs, including discrimination, exclusion, toxicity, information hazards, misinformation harms, malicious use, and human-computer interaction. However, there is limited research addressing these unethical issues of LLMs in higher education.
This paper aims to comprehensively examine the harmful behaviors displayed by GPT and its spin-offs in higher education. These behaviors encompass social and monolingual biases, issues of reliability and trustworthiness, data privacy and security concerns, toxicity in generated content, challenges in human-computer interaction, and environmental impact.

By conducting this study, valuable insights will be provided to educators, administrators, and students, enhancing their awareness of the ethical risks inherent in GPT and its various versions. This will enable them to proactively address and mitigate potential ethical concerns arising from the use of LLMs in educational settings.

**Social Biases and Monolingual biases**

Prior to delving into social biases in higher education, it is essential to define the concept of fairness. There is no consensus on the definition of fairness among scientists from various disciplines such as psychology, philosophy, and computer science (18). However, the broad definition of fairness is the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics (19). Thus, the notion of bias refers to injustice and unfairness that is tilted toward a specific segment of society.

GPT and its variants (i.e., GPT2, GPT3, GPT3.5, GPT4, etc.) are trained on massive digitized text sources such as digital books, online articles, historical archives, and more. The models obtained from this training data are prone to various kinds of biases, such as population, social, behavioral, temporal, and content production biases, as outlined by Mehrabi et al. (19), although these biases are not categorized into LLMs. Training data is vulnerable to inherent biases because it is generated by human agents (19). In this sense, downstream uses of LLMs can indicate representational bias when the obtained model marginalizes or discriminates against certain groups of people (e.g., marginalized segments of society). As a result, LLMs may reflect the epistemology and ontology of a particular group (20).

For example, in the context of GPT’s popularity, administrators of different institutions might utilize GPTs to streamline the admission process (2). However, deploying LLMs (i.e., GPTs) to analyze resumes and other admission documents could lead to the rejection of discriminated groups (e.g., international applicants) or unjust recommendations for low-paying scholarships or assistantships for marginalized students (e.g., international students). To mitigate such bias, administrators must ensure that the training data is representative of the entire population and not tilted toward a specific group (17).

Furthermore, another concern for instructors is addressing monolingual features that lead to biases. Monolingual biases reflect the norms of the dominant culture within society (e.g., white English-speaking communities). As a result, variants of English-speaking languages and other languages might not be adequately represented within the obtained model (21). Models often struggle to perform well in low-resource languages, thereby creating an inequitable situation where minority groups lack the same advantages provided by LLMs (22, 23). Zhuo et al. propose that different types of languages and cultures must be incorporated into the training data to mitigate monolingual biases (17).

To examine this issue, the researcher asked GPT4 to explain the role of AI-based technology in education in two distinct languages: Pashto (spoken in Afghanistan) and standard English. The findings revealed a considerable disparity in performance between ChatGPT-4’s outputs in Pashto compared to the well-supported English language. Both the quality and quantity of the explanations produced in Pashto were significantly lower. Students from minority groups (i.e., international students) could not effectively and efficiently use such software compared to native speaker students, as they could obtain a robust and nuanced understanding of the concept in their mother tongue.

Figure 1. Presenting GPT4’s Response to an English Inquiry.
Reliability and Trustworthiness

Reliability in the context of language models (LMs) refers to the presence of misinformation and outdated information (17). LMs have the capacity to generate inaccurate, deceptive, nonsensical, and low-quality information, which can result from training on unreliable data. When models are exposed to false or misleading
information during training, inexperienced users, like students, may unknowingly rely on unreliable data. As a consequence, users who are not vigilant about unnoticed errors may be misled and develop misconceptions when relying heavily on seemingly factual information and data (24).

Another concern that can undermine reliability and trustworthiness within LMs is the issue of hallucination, where misleading and inaccurate information is generated. Hallucination can be categorized into intrinsic and extrinsic hallucination (25). Intrinsic hallucination occurs when the generated information conflicts with the existing source, while extrinsic hallucination arises when the generated information cannot be verified by available sources. For instance, hallucination is evident in the example below, where GPT4 furnishes students with erroneous information, leading learners to acquire false knowledge. Teachers should educate their students about the limitations of relying solely on information obtained from LLMs like GPT and its spin-offs.

Figure 3. A GPT4-Generated Instance of Misinformation

To address this issue, several methods are proposed, including alignment and optimizing tool usage (25). By aligning the generated information with reliable sources and optimizing the usage of LMs, users can enhance the reliability and trustworthiness of the information derived from these models. It is crucial to be cautious and critical when using LLMs, recognizing their potential for generating unreliable content, and employing verification and cross-referencing with reputable sources to ensure accuracy and validity.

Data Privacy and Security

LLMs pose security risks due to their inherent generality, as they are not specifically fine-tuned for a particular application. This makes them vulnerable to attacks that can potentially impact the applications that utilize them (5).
Another security concern related to LLMs is data leakage, which can expose sensitive information and compromise individual privacy and organizational security. To address data leakage, several methods can be employed, including differential privacy, which is a privacy-preserving technique used to protect sensitive information in datasets, and model distillation, which transfers knowledge from a complex model to a simpler, smaller one (17). In response to privacy issues, the government of Italy prohibited the use of GPT by its citizens in April due to concerns about the model collecting personal information and user preferences during interactions.

A thought-provoking question arises regarding how we can mitigate privacy and security concerns while utilizing LLMs in educational settings. One solution is to increase the awareness of IT staff at universities regarding data leakage, particularly in regard to sensitive information such as students’ medical histories, family backgrounds, and academic records. This heightened awareness can help reduce the occurrence of data breaches when using LLMs.

By adopting privacy-preserving techniques and taking proactive measures to safeguard sensitive data, educational institutions can make responsible and ethical use of LLMs while minimizing potential risks to privacy and security.

Toxicity

LLMs possess the ability to generate offensive content, including profanities, insults, and threats, as they often learn from prevalent online sources (17, 20). It is essential to acknowledge that LLMs have the potential to produce harmful and hateful language. For instance, the following GPT example shows how it could potentially incite teenagers to engage in violent behaviors. To address this issue, engineers and IT personnel must take measures to ensure that the training model is free of harmful and offensive language. They can actively work on refining the training data by detecting and excluding any hurtful languages (17).

As demonstrated in the example below, students can easily prompt GPT to provide them with a list of online gun stores, enabling them to make purchases. In the educational context, it is crucial for educators and IT staff to collaborate continuously to customize chatbots for students, ensuring that they are vigilant in preventing any improper use of the model. By actively monitoring and moderating the interactions with LLMs, educators can create a safer and more responsible environment for students.

Proactive measures need to be taken to prevent the generation of offensive content by LLMs. Engineers and IT personnel play a significant role in refining the training data and implementing safeguards to promote responsible and ethical use of language models in educational settings.

Figure 4. An Example of Offensive and Toxic Content Generated by GPT
Human-Computer Interaction

The primary application of LM is to engage human users in a conversational manner, effectively serving as conversational agents (26). These language models can be integrated into dialogue-based systems, creating a simulated environment where users interact with them as if they were conversing with humans (27). However, notable incidents have highlighted the potential emotional attachment that users can develop towards chatbots.

For instance, there was a case involving Black Lemoine, who resigned from Google on June 14, 2022, due to his strong emotional connection to a chatbot he had developed and interacted with as if it were human. Additionally, on March 23, 2023, a young Belgian man tragically passed away after becoming deeply immersed in conversations with a chatbot named ELIZA (28).

While GPT and its various versions can be considered as peer tutors for students, assisting them in finding solutions and submitting assignments, continuous interaction with AI agents can lead to potential safety concerns. Students might perceive chatbot-like GPTs as human-like, which could result in them becoming overly dependent or emotionally attached to the AI.

To address these potential problems, it is essential for all students exposed to such software to receive proper education on how to use it and understand its limitations and capabilities. Educators and administrators must actively promote responsible use of AI agents and emphasize the importance of maintaining a balanced and healthy relationship with these language models.

Environmental Harm

The environmental impact of LLMs, as highlighted by Weidinger, et al. (16) and Borji, et al. (5), is primarily associated with their carbon footprint. The development and training of large language models require substantial computational resources, resulting in significant energy consumption. This energy-intensive process incurs costs related to hardware and electricity. Additionally, the environmental implications of training models are a concern due to the substantial energy requirements of modern tensor processing hardware, leading to carbon emissions (29).

Figure 5: Parameter Training Comparison among Different Large Language Models.
As indicated in the table above, the training model of GPT, which consists of one hundred seventy-five billion parameters, generates carbon emissions that surpass those produced by cars by more than five times (5). The size of large language models, such as GPT-4, becomes a crucial factor due to its detrimental impact on climate change.

Given the significant environmental consequences associated with LLMs, IT professionals must be mindful of their carbon footprint and explore strategies to mitigate it. Developing more energy-efficient training methods and utilizing renewable energy sources are essential steps in reducing the environmental impact of LLMs and promoting sustainable practices. By taking proactive measures, we can contribute to a greener and more environmentally friendly future.

**Conclusion and Future Work**

While the adoption of LLMs, including different versions of GPT, can enhance human performance, it is essential to recognize that these systems can exhibit unethical behaviors. The current study outlines the visible problems of such software, regardless of which version of GPT is being used. The researcher considered how these aforementioned problems within LLMs can manifest in higher education. By diving into the fundamental unethical issues of LLMs, educators and stakeholders can gain insight and work towards mitigating the harmful biases perpetuated by these models. It is incumbent upon educators, stakeholders, administrators, and IT staff to closely collaborate and use approaches and strategies to reduce such unethical and unjust practices of LLMs in the context of education.

However, this study has certain limitations that need to be addressed in future research. Firstly, it does not delve into comprehensive approaches for mitigating the ethical hazards posed by LLMs, which should be considered in subsequent studies. Further research should focus on identifying and implementing effective strategies to ensure the responsible and ethical use of LLMs in educational settings.

**Acknowledgement**

I would like to convey our deep appreciation to the anonymous peer reviewer who offered invaluable and constructive feedback during the article’s preparation. Additionally, GPT played a significant role in aiding the content editing process for this article.

**Conflict of interest**
I do not have any conflicts of interest to disclose.

References


