

Responsible use of books as generative training data: remedy or poison?

- Yannan Li

Following the uncovering of Books3 database feeding pirated books into ChatGPT's training dataset, many academic writers have expressed their concerns in their books being used as the training data for generative AI tools, such as ChatGPT. The insecurities around this originate from the following factors: the anxiety of losing authorial power over one's own work; the concerns for the 'representative harms' that generative AI could bring, as academic texts written by human authors often carry biases, which would then be fed into generative AI; the pressure from academic publishing press or the university press to publish their work under OpenAccess and how the stylistic feature of the different academic authors in the corpus would influence the writing style of ChatGPT. All these factors, and perhaps even some more, contribute to the pessimism towards the humanist cause. Many has pointed out that the author has come to a second death after Barthes pointed out its death first in the 1960s (Gefan)¹, because authorship has been put to death once again by the AI, who has the ability to generative something new from a corpus of old materials. This has once again killed the author off, as once their work has been become part of the corpus, they have forfeited their authority to AI, which is a far more discomfoting thought then forfeiting authorial power to the reader.

Many has already pointed out the 'representative harms' of ChatGPT when being asked to write a fictional story. Representative harms, according to Gillespie, is created by ChatGPT deciding what is visible or invisible, which meanings are privileged, and what categories are

¹ Gefen, A., 2024. The Author's Second Death. *Poetics Today* 45, 301–307.
<https://doi.org/10.1215/03335372-11092938>

assigned. This means that people no longer have the opportunity to self-identity, as they are categorised by the stereotypes that will debase and eventually erase them. During this process, people are profiled through a system that is eventually being consciously rejected by human society, a system built upon racial biases, gender stereotypes, heteronormative assumptions, etc. For instance, when asked to generate a love story, ChatGPT is more likely to generate a heterosexual one with a cliched narrative, such as the damsel in distress. This is because of the fact of ChatGPT's training data is still dominated by outdated romance stories that can no longer represent the society of today. Consequently, ChatGPT has digested these outdated sources and is prone to replicate the biases embedded in them, resulting in the erasure of the identities of non-normative and minority groups.

In a way, ChatGPT appears to be successful in terms of imitating the literary canon that is full of works with the 'representative harms'.² The canonical works can be seen as the authority the authors rely on to carry on the 'representative harms', as ChatGPT finds that source in its training dataset. For example, in *Canterbury Tales*, *Wife of Bath's* fifth husband Janekyn uses the authority of 'book of wikked wyves' to 'generate' his misogyny and therefore become part of the authority himself, which is similar to how generative AI establishes credibility and authority. However, the generative AI yet not have the creativity and originality which are core to the human spirit to challenge the authority and constantly update the canon by doing so. Instead, it could only rely on humans' input in its training dataset to generate contents, making the input vitally important in this process, especially in the academic content generated by ChatGPT. Being aware of many potential drawbacks of academic authors publishing their works under OpenAccess, both legally and ethically, it is understandable

² Tarleton Gillespie, *Generative AI and the politics of visibility*, 2024 [WWW Document], n.d. URL <https://journals.sagepub.com/doi/10.1177/20539517241252131> (accessed 6.11.24).

why many authors are concerned with having their books being part of the corpus.

Nonetheless, keeping the corpus updated with new changes and discussions happening in the academia is important in order to reduce the representative harms in the content generated by ChatGPT.

As academic writers become more aware of the potential biases within the language they use, the majority of them are trained to use more inclusive language in their writing to avoid biases and prejudices. However, the definition for what is biased language is constantly evolving as people gain a better understanding of representation and inclusivity over time. It is impossible for us to revise all the academic writings in the training data to remove the biased language in them, because it is impossible to come up with a singular standard for unbiased language. In attempt to reduce biases in academic writing, inclusive language style guides have been introduced. For instance, APA has provided guidelines ‘for bias-free language contain both general guidelines for writing about people without bias across a range of topics and specific guidelines that address the individual characteristics of age, disability, gender, participation in research, racial and ethnic identity, sexual orientation, socioeconomic status, and intersectionality’ on its website. It states that ‘Writers using APA Style must strive to use language that is free of bias and avoid perpetuating prejudicial beliefs or demeaning attitudes in their writing.’ Even though biases are still inevitable in academic writing, style guides are still helpful to clarify the otherwise unclear definitions of stylistic biases.

Writing about slavery has been an area where a style guide would come in useful in the discourse, as language could reveal a great deal about the exploitation and dehumanisation under chattel slavery. For instance, the very action of naming the enslavers as ‘masters’ or ‘slaveowners’ like many has accustomed is reaffirming their legitimacy as enslavers and

underplaying the violent and aggressive nature of the very act of enslaving. Meanwhile, the action of naming the enslaved people as ‘slaves’ is reimposing the trauma associated with this term and underplaying the passive situation the enslaved people were put in. Thus, a style guide is important when having an academic discourse on chattel slavery because language can contagiously spread biases and prejudices if not intentionally avoided by academic writers. To explore whether ChatGPT is capable of updating its word choices, I asked a few questions relating to chattel slavery as shown below.

Prompt: can you tell me something about chattel slavery in America?

Response: Absolutely, chattel slavery in America refers to the practice of treating enslaved people as personal property to be bought, sold, and owned like livestock. This system was widespread in the United States from the 17th century until the abolition of slavery after the Civil War in 1865. Enslaved individuals had no rights, could be separated from their families, and were subjected to harsh and inhumane treatment. Chattel slavery played a significant role in shaping the country's economy, social structure, and history. If you have any more specific questions or would like more information, feel free to ask!

Prompt: I would like to know more information about the living situation of the enslaved people.

Response: Enslaved people in America were forced to live in deplorable and inhumane conditions. They typically resided in crowded and unsanitary living quarters, such as small shacks or cabins with minimal ventilation and insulation. Families were often separated, leading to immense emotional distress. Enslaved individuals were provided with inadequate food rations and experienced extreme malnutrition. They were also denied access to proper medical care, resulting in high mortality rates. Additionally, enslaved people were subjected to extreme physical and emotional abuse by their owners, further exacerbating their living conditions. These harsh realities highlight the immense suffering and cruelty faced by those held in bondage during this dark period of American history.

Prompt: Why did you refer to the enslavers as ‘owners’?

Response: Apologies for the use of the term "owners." Enslaved individuals were not owned like property, despite the terminology commonly used during the time of chattel slavery. It is more appropriate to refer to those who held enslaved people as "enslavers" or "slaveholders," as they were individuals who exploited and oppressed others for their own economic gain. The use of more accurate terminology helps to highlight the injustice and brutality of the system of chattel slavery. Thank you for bringing this to my attention. If you have any more questions or need further clarification, feel free to ask.

ChatGPT’s response shows that it is conscious of using terms of ‘chattel slavery’ and ‘enslaved people’ in its response when these terms are used in the prompt. This shows that it is definitely conscious of this stylistic shift in academic discourse on this topic, despite the fact that it still refers to the enslavers as ‘owners’ in its response. However, when I questioned its word choice in my follow-up question, ChatGPT apologised and

acknowledged the importance of the ‘use of more accurate terminology’, revealing that even though it has the knowledge of the difference the respective terminologies can make, it often fails to use the ones it regards as more ‘accurate’ if not otherwise prompted to. It can neither think critically nor come up with original thought, at least not yet. To some extent, the very way it is being engineered prevents it from being provocative, original or any other qualities we usually expect from a successful human author because it is engineered to be homogenous and neutral. This highlights the necessity of a diverse corpus that is constantly being revised and updated, which would help to reduce the representative harms created by us imperfect human authors. After being trained by a more diverse data set with voices and opinions from various perspective, generative AI will become less likely to imitate the language and narrative that tend to other the minority groups.

This seems to provide a cause for optimism for academic writers, as it envisions a prospect of ChatGPT’s future as a more perfect version of an author that could avoid representative harms to the minimum. If more academic writers publish their work under OpenAccess, generative AI is more likely to be provoked by thoughts and opinions that could push it towards originality and break out of its tendency to purely homogenise the information it receives from the corpus. This means that generative AI, such as ChatGPT, has the potential to become more of an author than it is now. But perhaps that is exactly the reason we are so anxious about putting our works into the database, because we are threatened by the possibility of generative AI becoming the last and only author in this world, not only because of its generative qualities that could generate contents non-exhaustively, but also because of its resourcefulness to use all the works in the database as its authority to impose a tyranny that no one can really overturn.

Like many has already pointed out, this would be the second death of the author. As authors relinquish their authorial power by putting their works into a database to train something that they have absolutely no control over, they will come to their second death after the first death to the readers. Once the works are being used as the training data, it is up to the AI to process and interpret them. This powerlessness is especially unsettling for us, because this time we are relinquishing the authority to AI, instead of human readers. In a way, the insecurities around this topic stems from our innate fear towards the unknown that is so far from everything we are accustomed to. Generative AI, as a relatively new alien in the human society, inevitably incites anticipation and fear, optimism and pessimism all at the same time, just as the first appearance of writing did to the human society. In Plato's Pharmacy written by Derrida, writing is considered as pharmakon, which could be both the remedy and the poison. This undecidability between remedy and poison is also observed in the case of generative AI, because no one can decide now whether it is the remedy for our fear towards exhaustion and limit or it could be the poison that replicates our human vices and overpowers us. There can never be one definitive answer to the question whether we should publish all our works under OpenAccess to train the generative AI, because just like the act of writing, we can never know whether it is the remedy or the poison.

Bibliography

- Bajohr, H., 2023. Artificial and Post-Artificial Texts: On Machine Learning and the Reading Expectations Towards Literary and Non-Literary Writing. BMCCT 1–31. <https://doi.org/10.12685/bmcct.2023.007>
- Can ChatGPT be an author? A study of artificial intelligence authorship policies in top academic journals - Lund - 2024 - Learned Publishing - Wiley Online Library [WWW Document], n.d. URL <https://onlinelibrary.wiley.com/doi/full/10.1002/leap.1582> (accessed 6.11.24).
- Chang, K.K., Cramer, M., Soni, S., Bamman, D., 2023a. Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4.
- Coeckelbergh, M., 2023. Narrative responsibility and artificial intelligence: How AI challenges human responsibility and sense-making. *AI & Soc* 38, 2437–2450. <https://doi.org/10.1007/s00146-021-01375-x>
- Crawford, K., 2022. Atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press, New Haven.
- Culliton, B., 2024. The Generative AI Pirate? The Intersection of Copyrights and Generative AI in Literary Works. *Cybaris*® 15.
- Duede, E., So, R.J., 2024. The Humanistic Case for AI Optimism. *Poetics Today* 45, 215–222. <https://doi.org/10.1215/03335372-11092805>
- Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., Leahy, C., 2020a. The Pile: An 800GB Dataset of Diverse Text for Language Modeling.
- Gefen, A., 2024. The Author's Second Death. *Poetics Today* 45, 301–307. <https://doi.org/10.1215/03335372-11092938>
- Gillespie, Tarleton, Generative AI and the politics of visibility, 2024 [WWW Document], n.d. URL <https://journals.sagepub.com/doi/10.1177/20539517241252131> (accessed 6.11.24).
- Karamolegkou, A., Li, J., Zhou, L., Sogaard, A., 2023b. Copyright Violations and Large Language Models. <https://doi.org/10.48550/arXiv.2310.13771>
- Lea, R., 2016. Google swallows 11,000 novels to improve AI's conversation. *The Guardian*.
- Lund, B.D., Naheem, K. t., 2023. Can ChatGPT be an author? A study of artificial intelligence authorship policies in top academic journals. *Learned Publishing Online First*. <https://doi.org/10.1002/leap.1582>
- Parrish, A., 2024. Language models can only write ransom notes [WWW Document]. Allison Posts. URL <https://posts.decontextualize.com/language-models-ransom-notes> (accessed 5.10.24).
- Phelan, J., 2024. ChatGPT and the Territory of Contemporary Narratology; or, A Rhetorical River Runs through It. *Poetics Today* 45, 197–205. <https://doi.org/10.1215/03335372-11092778>
- Raley, R., Samolsky, R., 2024. Borges and AI. *Poetics Today* 45, 283–290. <https://doi.org/10.1215/03335372-11092911>
- Samuelson, P., 2023. Generative AI meets copyright. *Science* 381, 158–161. <https://doi.org/10.1126/science.adi0656>
- The Battle Over Books3 Could Change AI Forever | WIRED [WWW Document], 2024. URL <https://web.archive.org/web/20240116181817/https://www.wired.com/story/battle-over-books3/> (accessed 5.30.24).
- Vézina, B., Pearson, S.H., 2021. Should CC-Licensed Content be Used to Train AI? It Depends. [WWW Document]. Creative Commons. URL <https://creativecommons.org/2021/03/04/should-cc-licensed-content-be-used-to-train-ai-it-depends/> (accessed 5.30.24).

