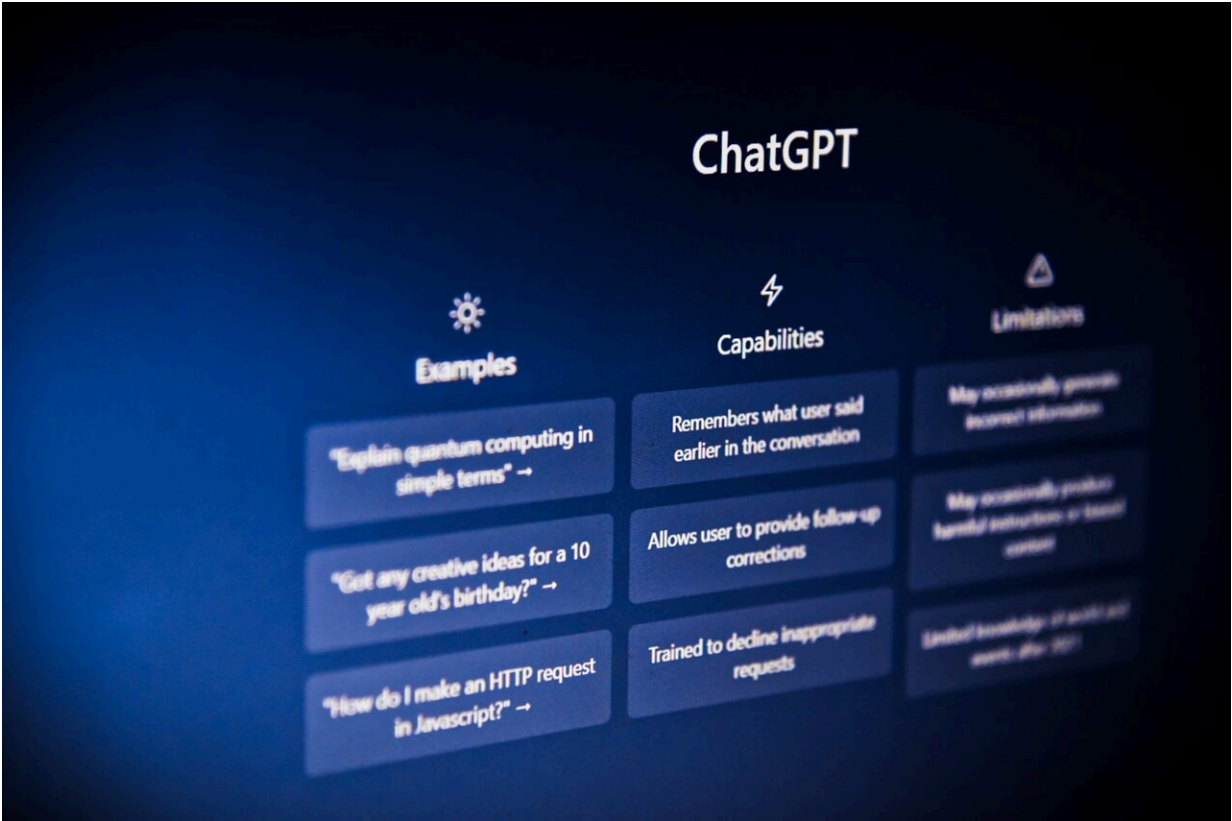# Ethical, legal issues raised by ChatGPT training literature

May 8 2023, by Peter Grad

Credit: Unsplash/CC0 Public Domain

Researchers at the University of California, Berkeley, say ChatGPT has memorized a large number of copyrighted works and that inclusion of such data can introduce bias to analytics conducted with OpenAI models.

Berkeley's Kent Chang, Mackenzie Cramer, Sandeep Son and David Bamman reported their findings on April 28 in a paper on the *arXiv* preprint server titled, "Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4."

While the revelation immediately raises questions of propriety and copyright protections, the researchers' primary interests are in transparency and the potential for unseen biases when those relying on OpenAI remain in the dark about what sources were included, and excluded, from input.

"We find that OpenAI models have memorized a wide collection of copyrighted materials, and that the degree of memorization is tied to the frequency with which passages of those books appear on the web," the researchers said.

"The ability of these models to memorize an unknown set of books complicates assessments of measurement validity for cultural analytics by contaminating test data," they cautioned.

For instance, the researchers noted that [science fiction](link) and fantasy books dominate the list of memorized books, presenting a built-in bias on the nature of responses ChatGPT may provide.

"The accuracy of such models is strongly dependent on the frequency with which a model has seen information in the training data, calling into question their ability to generalize," they said. Such models "present a challenge" when it comes to validating results since few if any details about data used to train the models are known to the public.

"Knowing what books a model has been trained on is critical to assess such sources of bias," they said.

"Our work here has shown that OpenAI models know about books in proportion to their popularity on the web."

Works detected in the Berkeley study include "Harry Potter," "1984," "Lord of the Rings," "Hunger Games," "Hitchhiker's Guide to the Galaxy," "Fahrenheit 451," "A Game of Thrones" and "Dune."

While ChatGPT was found to be quite knowledgeable about works in the public domain, lesser known works such as Global Anglophone Literature—readings aimed beyond core English-speaking nations that include Africa, Asia and the Caribbean—were largely unknown. Also overlooked were works from the Black Book Interactive Project and Black Caucus Library Association award winners.

"We should be thinking about whose narrative experiences are encoded in these models, and how that influences other behaviors," Bamman, one of the Berkeley researchers, said in a recent Tweet. He added, "popular texts are probably not good barometers of model performance [given] the bias toward sci-fi/fantasy."

The researchers said their findings make the case for the use of open models that disclose training data.

Meanwhile, major legal challenges are likely in the near future. What are the limitations of "fair use" when copying text? Who owns the copyright on text generated in full or in part by ChatGPT? Who prevails when copyright protection is sought for multiple similar or identical outputs by multiple parties?

And perhaps a more interesting question: Is machine language copyrightable all?

Some may recall the famous "Macaque selfie" case in which a monkey

snapped photos of itself with equipment left behind by a professional photographer. The photographer sued publications that used the fascinating photos, but they argued that since the photographer did not take the photos he could not claim copyright protection. PETA argued the monkey should hold the copyright.

Years of legal battles led to a 2018 ruling that affirmed non-humans have no authority to claim copyright.

Will that extend to ChatGPT literature?

More information: Kent K. Chang et al, Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4, *arXiv* (2023). [DOI: 10.48550/arxiv.2305.00118](#)

© 2023 Science X Network