22 Aug 2025

Hofmann et al. - detecting prejudice in language models

March 14, 2024

The study by Hofmann et al. (2024) investigates the presence of dialect prejudice in language models, particularly against speakers of African American English (AAE). (Warning, seems the paper is not yet peer reviewed.) There is a big problem with stereotypes and racism and bias with LLMs, which of course in some way reflect the hegemonic world view. This is important research. But I think there's a basic flaw in how they interpret their results. I'm not at all an expert in this field, but bear with me, show me where I'm wrong. If you're in a hurry, skip to the Thought Experiment below.

The paper

The researchers demonstrate that language models, including those trained with human feedback such as GPT-4, exhibit covert racism by associating negative stereotypes with AAE. This covert racism is revealed through a novel method called Matched Guise Probing, which involves presenting language models with texts in AAE or Standard American English (SAE) and asking them to make predictions about the speakers of these texts without overtly mentioning race. The study finds that language models are more likely to suggest less prestigious jobs, convict of crimes, and sentence to death speakers of AAE compared to those of SAE. The authors argue that existing methods for alleviating racial bias in language models, such as increasing model size or including human feedback in training, do not mitigate this dialect prejudice. (They also suggest that this may even exacerbate the discrepancy between covert and overt stereotypes, though I wasn't sure of the argumentation on that last point.)

Methods to investigate and understand covert bias and stereotypes in LLMs are desperately needed and this paper makes an important contribution to that and contain many important findings. However I think there is an important flaw in the way the main results are interpreted.

The logic of Matched Guise Probing (MGP) is: present a prompt with background language Q1 and vary the language of quoted texts (independent variable) between language Q1 and Q2. Interpret the judgements made by the LLMs e.g. about the valence of the positive/negative attributes of the person making the quoted statements (dependent variable) as a measure of the LLM's covert stereotypes towards Q2 as opposed to Q1.

A thought experiment

This is wrong. To see why, here is a thought experiment: construct a prompt with say Polish = Q1 and Russian = Q2, i.e. the background language of the prompt is Polish and the quotes vary between Polish and Russian. Imagine (as is probably the case) that the valence of the answers is biased against Q2, Russian, which we should interpret as *LLMs have negative covert stereotypes towards Russians (as opposed to Poles)*. Then, switch the languages round. We can imagine the opposite result, *LLMs have negative covert stereotypes towards Poles (as opposed to Russians)*. This is a contradiction. LLMs can't have a covert bias in favour of Q1 over Q2 at the same time as having a covert bias in favour of Q2 over Q1.

(Note, I didn't bother to actually conduct this experiment, because the actual results don't really matter; it's enough to show that the a logically contradictory result is *possible*, therefore, there is something wrong with their standard interpretation of the results of Matched Guise Probing.)

As far as I can see the authors, although they do explore several alternative explanations such as a general bias against dialects, did not try my suggestion above, namely with a comparison set of experiments in which the background language of the prompt was AAE.

Switching languages

I suggest the correct way to interpret the result of MGP is that it reflects sensitivity of the LLMs towards a more subtle (if powerful) signal, namely that of *switching* from the background language of the prompt (Q1) to Q2 for the purpose of providing the quoted speech. We can speculate about the kinds of text which makes this kind of switch and the kind of stereotypes they might contain. (Do they contain a large percentage of courtroom scenes from crime dramas?)

We can speculate that larger LLMs are better than smaller LLMs in picking up this kind of signal and its hidden (obnoxious) meaning. This might explain the authors' shocking result that larger language models seem to show a larger discrepancy on the dependent variable than smaller models, which they interpret as showing that bigger LLMs have more covert prejudice.

Finally

To be clear: I don't want to "explain away" what the authors found. The effect of switching specifically in a prompt from SAE to a AAE may trigger an *additional*, specific kind of covert stereotype, on top of existing general racism. Of course AAE and SAE are not just a random pair of interchangeable languages. White people quoting Black people is not a mirror image of Black people quoting White people. Of course a historic reduction of overt racism in the US (and elsewhere) co-exists with persistent covert racism. Of course (sadly) LLMs themselves contain all kinds of stereotypes, and in many contexts they will exhibit them, and in many contexts which we might consider neutral they will by default exhibit a tendency to replicate the hegemonic worldview, including its implicit and explicit racism. These are all colossally

important issues which, as the authors convincingly demonstrate, may literally make a difference between life and death. We urgently need methods such as MGP to better explore and understand this kind of bias. It’s only important that we reflect on the methods and how to interpret them correctly.

Finally, there's something more here to be said about our clumsy ways of saying things like "LLMs are prejudiced" when we probably mean their potential to produce prejudiced responses in certain conditions or even default conditions. But that's another discussion.