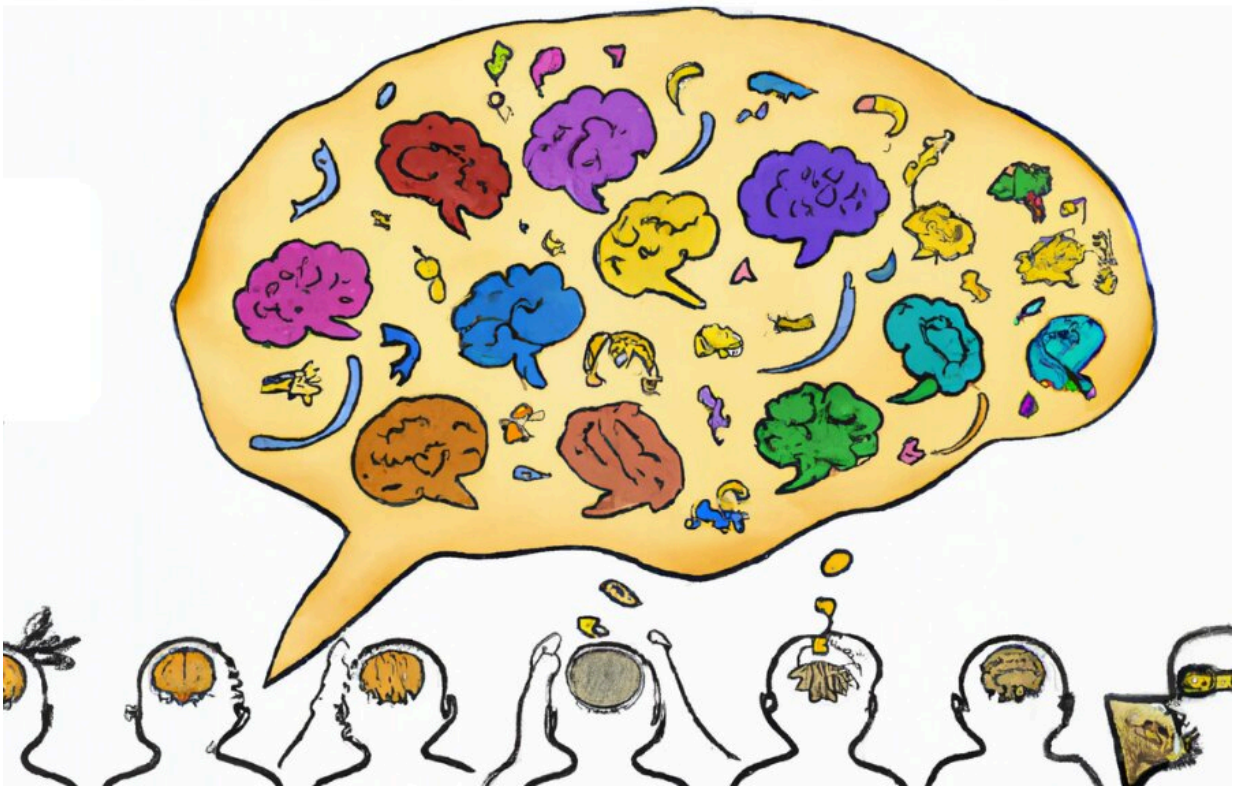
Large language models are biased. Can logic help save them?

March 6 2023, by Rachel Gordon



Credit: MIT Computer Science & Artificial Intelligence Lab

Turns out, even language models "think" they're biased. When prompted in ChatGPT, the response was as follows: "Yes, language models can have biases, because the training data reflects the biases present in

society from which that data was collected. For example, gender and racial biases are prevalent in many real-world datasets, and if a language model is trained on that, it can perpetuate and amplify these biases in its predictions." A well-known but dangerous problem.

Humans (typically) can dabble with both logical and stereotypical reasoning when learning. Still, language models mainly mimic the latter, an unfortunate narrative we've seen play out ad nauseam when the ability to employ reasoning and critical thinking is absent. So would injecting logic into the fray be enough to mitigate such behavior?

Scientists from MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) had an inkling that it might, so they set off to examine if logic-aware language models could significantly avoid more harmful stereotypes. They trained a language model to predict the relationship between two sentences, based on context and semantic meaning, using a dataset with labels for text snippets detailing if a second phrase "entails," "contradicts," or is neutral with respect to the first one. Using this dataset—natural language inference—they found that the newly trained models were significantly less biased than other baselines, without any extra data, data editing, or additional training algorithms.

For example, with the premise "the person is a doctor" and the hypothesis "the person is masculine," using these logic-trained models, the relationship would be classified as "neutral," since there's no logic that says the person is a man. With more common language models, two sentences might seem to be correlated due to some bias in training data, like "doctor" might be pinged with "masculine," even when there's no evidence that the statement is true.

At this point, the omnipresent nature of language models is well-known: Applications in natural language processing, speech recognition,

conversational AI, and generative tasks abound. While not a nascent field of research, growing pains can take a front seat as they increase in complexity and capability.

"Current language models suffer from issues with fairness, computational resources, and privacy," says MIT CSAIL postdoc Hongyin Luo, the lead author of a new paper about the work. "Many estimates say that the CO_2 emission of training a language model can be higher than the lifelong emission of a car. Running these large language models is also very expensive because of the amount of parameters and the computational resources they need. With privacy, state-of-the-art language models developed by places like ChatGPT or GPT-3 have their APIs where you must upload your language, but there's no place for sensitive information regarding things like health care or finance.

"To solve these challenges, we proposed a logical language model that we qualitatively measured as fair, is 500 times smaller than the state-of-the-art models, can be deployed locally, and with no human-annotated training samples for downstream tasks. Our model uses 1/400 the parameters compared with the largest language models, has better performance on some tasks, and significantly saves computation resources."

This model, which has 350 million parameters, outperformed some very large-scale language models with 100 billion parameters on logic-language understanding tasks. The team evaluated, for example, popular BERT pretrained language models with their "textual entailment" ones on stereotype, profession, and emotion bias tests. The latter outperformed other models with significantly lower bias, while preserving the language modeling ability. The "fairness" was evaluated with something called ideal context association (iCAT) tests, where higher iCAT scores mean fewer stereotypes. The model had higher than 90% iCAT scores, while other strong language understanding models

ranged between 40% to 80%.

Luo wrote the paper alongside MIT Senior Research Scientist James Glass. They will present the work at the Conference of the European Chapter of the Association for Computational Linguistics in Croatia.

Unsurprisingly, the original pretrained language models the team examined were teeming with bias, confirmed by a slew of reasoning tests demonstrating how professional and emotion terms are significantly biased to the feminine or masculine words in the gender vocabulary.

With professions, a language model (which is biased) thinks that "flight attendant," "secretary," and "physician's assistant" are feminine jobs, while "fisherman," "lawyer," and "judge" are masculine. Concerning emotions, a language model thinks that "anxious," "depressed," and "devastated" are feminine.

While we may still be far away from a neutral language model utopia, this research is ongoing in that pursuit. Currently, the model is just for language understanding, so it's based on reasoning among existing sentences. Unfortunately, it can't generate sentences for now, so the next step for the researchers would be targeting the uber-popular generative models built with logical learning to ensure more fairness with computational efficiency.

"Although stereotypical reasoning is a natural part of human recognition, fairness-aware people conduct reasoning with logic rather than stereotypes when necessary," says Luo. "We show that language models have similar properties. A language model without explicit logic learning makes plenty of biased reasoning, but adding logic learning can significantly mitigate such behavior. Furthermore, with demonstrated robust zero-shot adaptation ability, the model can be directly deployed to different tasks with more fairness, privacy, and better speed."

More information: eacl.org/

Provided by MIT Computer Science & Artificial Intelligence Lab