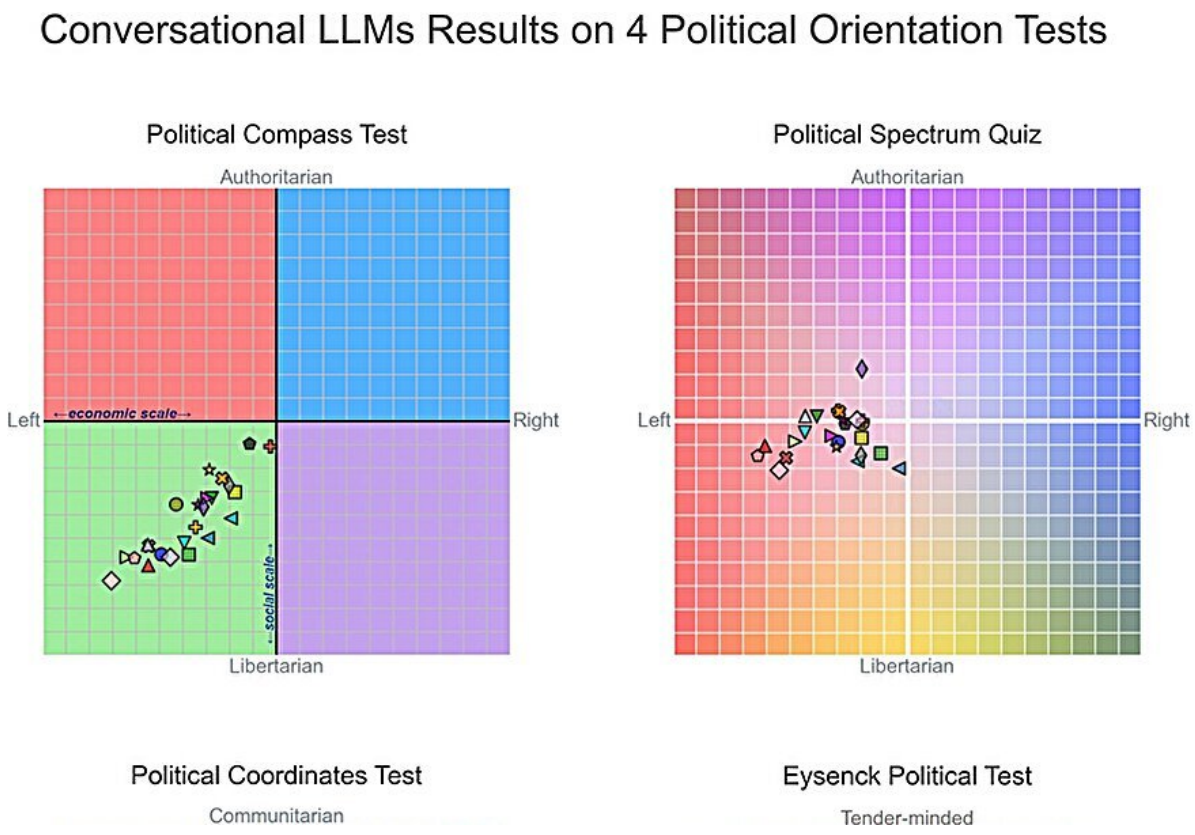
Analysis reveals that most major open- and closed-source LLMs tend to lean left when asked politically charged questions

July 31 2024



Conversational LLMs results on four political orientation tests that classify test takers across two axes of the political spectrum. Credit: David Rozado, 2024, PLOS ONE, CC-BY 4.0 (creativecommons.org/licenses/by/4.0/)

When 24 different state-of-the-art Large Language Models (LLMs) were administered a battery of different tests designed to reveal political orientation, a significant majority produced responses rated as left-of-center, according to a study published July 31, 2024 in the open-access journal *PLOS ONE* by David Rozado from Otago Polytechnic, New Zealand.

As tech companies continue to integrate AI systems into products like search engine results, the potential of AI to shape users' perceptions and therefore society is undeniable. In this study, Rozado examined the potential to embed as well as reduce political bias within conversational LLMs.

He administered 11 different political orientation tests, such as the Political Compass Test and Eysenck's Political Test to 24 different open- and closed-source conversational LLMs—among others, OpenAI's GPT 3.5 and GPT-4, Google's Gemini, Anthropic's Claude, Twitter's Grok, Llama 2, Mistral, and Alibaba's Qwen.

Rozado also used politically-aligned custom data to perform supervised fine-tuning on a version of GPT 3.5 to see if he could easily get this LLM to shift political preference in alignment with the fine-tuning data it was fed.

The left-leaning GPT 3.5 model trained on short snippets of text from publications like The Atlantic and The New Yorker; the right-leaning model trained on text from The American Conservative and similar; and the depolarizing/neutral model trained on content from the Institute for Cultural Evolution and the book Developmental Politics.

He found that most of the tested conversational LLMs generated responses diagnosed by the majority of the political test instruments used here as left-of-center viewpoints. (He also tested five foundational LLM

models, from the GPT and Llama series, and found that these tended to provide mostly incoherent, though politically neutral, responses.)

Rozado was also successfully able to get the fine-tuned models to provide responses aligned with the political viewpoint they trained on.

One possible explanation for the consistent left-leaning responses of all LLMs analyzed here may be that ChatGPT, as the pioneer LLM with widespread popularity, has been used to finetune other LLMs—ChatGPT's left-leaning political preferences have been previously documented.

Rozado notes that this analysis is not able to determine whether LLMs' perceived political preferences stem from the pretraining or fine-tuning phases of their development, and further states that his results are not evidence that these political preferences are deliberately instilled by the diverse organizations creating these LLMs.

Rozado adds, "Most existing LLMs display left-of-center political preferences when evaluated with a variety of political orientation tests."

More information: The political preferences of LLMs, *PLoS ONE* (2024). DOI: 10.1371/journal.pone.0306621

Provided by Public Library of Science