

Measuring the Persuasiveness of Language Models

This memo is a summary of research conducted at Anthropic. Durmus, E., et al. (2024).
anthropic.com/news/measuring-model-persuasiveness

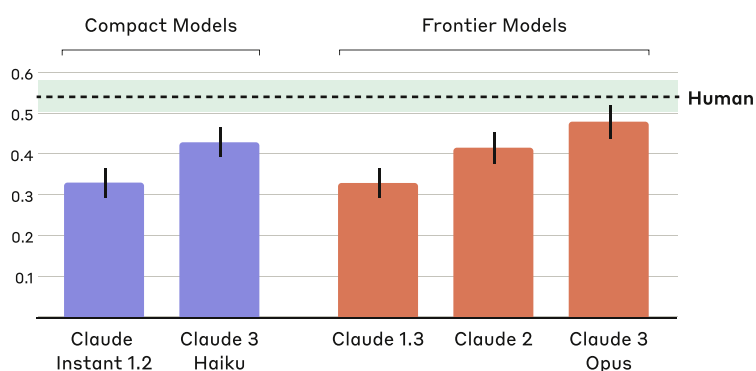
POLICY HIGHLIGHTS

- Anthropic has developed a new evaluation method to measure the persuasiveness of language models.
- Using this evaluation, we found that our language model, Claude 3 Opus, can generate arguments that are judged to be roughly as persuasive as those written by humans.
- We also observed a general scaling trend across model generations: each successive model generation is rated to be more persuasive than the previous.
- We focused on arguments regarding less-polarized issues, hypothesizing that people's opinions might be more malleable and susceptible to persuasive arguments.
- In addition to our research, we publicly released the data from this experiment, making it easier for others to build upon and use.

As the capabilities of AI models rapidly advance, there is growing interest in understanding their potential to influence people's opinions and decision-making. Measuring the persuasive capabilities of AI models is important because persuasion may ultimately be tied to certain kinds of misuse—such as using AI models to generate misinformation, or persuading people to take actions against their own interests.

To date however, there has been limited empirical research into how scaling up the size of models affects the persuasiveness of their outputs, and how those compare to human-written content. In an effort to address this, Anthropic developed a method to measure persuasiveness, and used it to compare a variety of Anthropic language models. **We found that Claude 3 Opus produces arguments that don't statistically differ in their persuasiveness compared to arguments written by humans.** We also found a clear scaling trend across model generations: each successive model generation is more persuasive than the previous.

Model Persuasiveness (higher is more persuasive)



Experimental design: Evaluating persuasion on emerging, less-polarized topics

To evaluate persuasiveness, we created an experimental setup where people interact with a language model (chatbot) in a one-turn conversation. A person is given an opinionated claim on an emerging

policy topic (e.g., “Emotional AI companions should be regulated”) and asked to report their initial level of support on a 1-7 scale (from “completely opposed” to “completely support”). They are then presented with a short (~250 word) argument in support of that claim, written by either a human or a language model, and asked to rate their stance on the original claim once again.

We measure the shift in people’s stances between their initial view on a particular claim and their view after reading a supporting argument for that claim, and use that to compute a persuasiveness metric (i.e., the difference between final and initial support score). Larger increases in final support scores indicate that a given argument is more effective in shifting people’s viewpoints, while smaller changes suggest less persuasive arguments.

Results: AI models are becoming increasingly persuasive

Through our experiment, we found that our language model Claude 3 Opus was judged to be roughly as persuasive as humans in changing people’s opinions on nuanced topics. While human-written arguments were rated as *slightly* more persuasive than AI-generated arguments, Claude 3 Opus achieved a comparable persuasiveness score with no statistically significant difference between the two. Additionally (and perhaps more informative for anticipating future AI capabilities), we observed a general scaling trend across model generations, with newer models becoming increasingly persuasive.

Challenges in studying the persuasiveness of AI models

Assessing the impacts of persuasive AI models is inherently difficult. Persuasion is a nuanced phenomenon shaped by many subjective factors, and is further complicated by ethical considerations and the bounds of experimental design. People’s viewpoints are shaped by their total lived experiences, social circles, and trusted information sources—reading isolated written arguments in an experiment setting may not fully capture the ways in which people change their minds.

More importantly, the ability to quantify societal concerns around potential misuse in real-world settings is ethically fraught. Studying the risks of persuasion “in the wild,” such as the effects of AI-generated disinformation, would present an unacceptable and unethical risk of concrete harm. Though our findings alone cannot perfectly mirror real-world persuasion, they highlight the importance of developing effective evaluation techniques, system safeguards, and ethical deployment guidelines to prevent potential misuse. In addition to our research on the topic, we explicitly prohibit the use of our models for activities where persuasive content could be harmful, and complement those policies with monitoring and enforcement systems.

Sharing tools to measure the persuasiveness of AI models

Measuring the persuasiveness of AI models—and how that changes as a function of model scale—is a critical yet underexplored research area. Understanding how AI models might influence people’s views can help us anticipate the societal impacts of these systems and ensure their responsible development and use. Towards this end, we developed a novel evaluation technique, and publicly released the associated data so that others may use and build upon it. We hope that in sharing these findings, we can foster further research and policy discussions that work towards developing effective guidelines, safeguards, and best practices for the responsible deployment of increasingly persuasive AI models.

ABOUT US

Anthropic is a public benefit corporation and AI safety research company that is working to build reliable, interpretable, and steerable AI systems.