Large Language Models Are More Persuasive Than Incentivized Human Persuaders

Schoenegger et al (2025)

Context

AI has advanced rapidly which has sparked concern over it's potential harm individuals and society. One key area of concern is regarding AI's ability to persuade individuals through personalised conversation to engage in behaviours they would have otherwise not have.

This study compares the persuasion (both truthful & deceptive) capabilities of a large language model (LLM) against incentivised human persuaders in an interactive real-time conversational quiz setting.

Key Insights

LLMs outperformed incentivised humans

The authors find that LLMs outperform incentivised humans in influencing participant responses in the quiz in both truthful and deceptive persuasion. LLM persuasion resulted in higher accuracy under truthful persuasion and lower accuracy under deceptive persuasion, so they are more effective in misleading participants than incentivised humans.

Ethical & regulatory governance is imperative

Since LLMs can persuade in both truthful and deceptive this creates a huge risk around misuse of LLMs particularly due to its easy scalability. The authors emphasise the important of regulatory guardrails, monitoring and safety mechanisms to prevent misleading users.

LLMs increase confidence, but their effect diminishes over time

Participants in the LLM persuasion condition reported higher confidence levels than participants in the other conditions. However, despite LLMs achieving stronger compliance than incentivised humans, their advantage decrease with each successive question which may be due to participants becoming used to the LLM's persuasion style over time.

Implications

- How are you using LLMs and is there potential for them to persuade towards certain undesirable behaviours or actions?
- How are you monitoring this? What safety mechanisms or guardrails do you have in place?
- How do you think your customers would react to interacting with an LLM?

Schoenegger, P., Salvi, F., Liu, J., Nan, X., & Debnath, R. (2025). Large language models are more persuasive than incentivized human persuaders. arXiv preprint arXiv:2505.09662. https://arxiv.org/abs/2505.09662