UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Detecting Incentive Skepticism in AI Persuasion Dialogues with LLM-Based Stance Inference

Permalink

https://escholarship.org/uc/item/5mj781qx

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Mori, Ryutaro

Horiguchi, Ilya

Publication Date

2026

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Detecting Incentive Skepticism in AI Persuasion Dialogues with LLM-Based Stance Inference

Ryutaro Mori
RIKEN, Japan

Ilya Horiguchi
The University of Tokyo, Japan

Abstract

Large language models (LLMs) enable tailoring of persuasive messages at scale, raising hopes and concerns that persuasion may become a problem of optimizing message content. However, people can resist persuasion for qualitatively different reasons, including doubts about the persuader's motives. Here, we reanalyze a public dataset of multi-turn human-LLM persuasion dialogues to examine whether such resistance can be detected early from recipients' own replies. We use an LLM as a proxy forward model of human response: we specify different stances in the system prompt and compute how likely the model is to generate the observed rebuttal. We find that skepticism about the sender's incentives, where recipients attend to possible incentive misalignment and discount message content, is common. Such skepticism predicts lower subsequent persuasion success and a higher risk of backfire. These findings highlight limits of content optimization and suggest the methodological potential of seeing LLMs as superpositions of human cognition.

Keywords: Human-AI interaction; persuasion; LLM

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* ©2026 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).