Explicit Cooperation Shapes Human-Like Multi-Agent LLM Negotiation

Yanru Jiang
University of California Los Angeles, Los Angeles, California, United States

Gülşah Akçakır
University of California, Los Angeles, Los Angeles, California, United States

Abstract

Humans develop cooperation heuristics in social decision-making, either intuitively or deliberatively. Large language models (LLMs), which exhibit human-like biases across cognitive domains, may acquire prosocial tendencies through instruction tuning, enabling cooperative behavior in strategic reasoning games. However, most studies of this kind either focus on cooperative language generation or explicitly instruct LLMs to cooperate, deviating from the inherent cooperation heuristics of humans. Using negotiation role-play simulations with BATNA (Best Alternative to a Negotiated Agreement), we found that LLMs struggle with cooperation in the absence of explicit instructions, leading to a 50–80% lower success rate than in instructed scenarios and 50–60% lower than human performance reported in past studies. Implicitly inducing cooperation through personality traits had inconsistent effects, with agreeableness showing marginal influence and other traits exhibiting no systematic impact. These findings suggest that personality-based cooperation cues are subtle, and explicit instructions remain essential for multi-agent LLMs to approximate human-like negotiation.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*