Project Abstract

Project members: Lucia Huo, Nicola Lawford

Modern AI systems have been deployed broadly as decision-makers across diverse domains, including healthcare, law, education, and finance. These AI systems predominantly operate in two perspectives for decision-making. The first is the **first-party perspective**, where the AI system substitutes or assists the user's decision-making. Second is the **third-party perspective**, where they are asked to operate in relation to another agent that assists the user in decision-making. Prior work (e.g. LLM-as-a-judge or self-reflection for model performance improvement) indicates that asking AI to take on a third-party perspective affords performance gains, though no work has explicitly examined the impact of a perspective shift from first- to third-party. Our work evaluates performance of various LLMs in medical QnA when the task is framed to the LLM in the first-party perspective versus third-party perspective. AI that operates better in the third-party perspective invites possibilities of AI as judges and self-auditors. Our project also explores - at a surface level - the paradigm of using psychological knowledge to advance AI research and governance.