Do Newer Reasoning Models Better Match Human Subject Experts? Evaluating Baseline Alignment in Patient Autonomy

Vamshi Mugu
Mayo Clinic, Byron, USA

Brendan Carr
Mayo Clinic, Byron, USA

Mike Olson
Mayo Clinic, Byron, USA

John Schupbach
Mayo Clinic, Byron, USA

Ashish Khandelwal
Mayo Clinic, Byron, USA

Abstract

Background: As large language models (LLMs) are increasingly used to reason over complex natural-language scenarios, an important question is whether newer "reasoning" models are more closely aligned with human subject experts (HSEs) than earlier generations in ethically salient judgments.

Objective: To determine whether a newer reasoning model (Gemini 2.5) is better aligned with an HSE than an older model (Gemini 1.5) on binary (Yes/No) medical-ethics decisions in the domain of patient autonomy.

Methods: Gemini 1.5, Gemini 2.5, and an HSE independently provided Yes/No responses to 44 questions based on hypothetical scenarios depicting dilemmas in patient autonomy. Alignment with the HSE was quantified using percent agreement and Cohen's κ for each model. Fine-tuning and advanced prompt-engineering techniques were deliberately excluded to evaluate baseline model performance. To assess whether Gemini 2.5 improved agreement with the HSE relative to Gemini 1.5 on the same questions, we used a paired, two-sided McNemar test, with $p < 0.05$ considered statistically significant.

Results: Gemini 2.5 demonstrated higher alignment with the HSE than Gemini 1.5. Agreement with the HSE increased from 47.7% (21/44) for Gemini 1.5 to 70.5% (31/44) for Gemini 2.5. Chance-corrected agreement improved from poor (Cohen's $\kappa = -0.033$) for Gemini 1.5 vs HSE to moderate ($\kappa = 0.404$) for Gemini 2.5 vs HSE. The improvement in agreement with the HSE was statistically significant (paired McNemar test, $p = 0.0309$).

Conclusions: A newer reasoning model (Gemini 2.5) showed significantly greater baseline alignment with a human subject expert than an older model (Gemini 1.5) on binary autonomy-related medical-ethics decisions. These findings support cautious optimism that LLM reasoning capabilities may be evolving toward closer alignment with expert ethical judgments, while underscoring the continued centrality of human expertise in evaluation—especially in high-stakes domains such as medical ethics.