Moral Choice and Collective Reasoning in Large Language Models

Amir Amangeldi, Natalie DellaMaria, Prakrit Baruah, Zaina Edelson

{aamangeldi, ndellamaria, pbaruah, zedelson}@g.harvard.edu

CS2881 Final Project

Motivation

As large language models transition from research artifacts to deployed agents making consequential decisions, understanding how they reason about ethical trade-offs and interact with each other becomes essential. Recent evidence reveals concerning variability in LLM moral reasoning. Different models exhibit dramatically different moral frameworks, and decisions are highly sensitive to contextual manipulation.

Key Questions

(1) How consistent are individual LLMs in moral choices, and how malleable to prompt engineering?

(2) When multiple LLMs deliberate, do they converge toward shared frameworks or develop power asymmetries?

(3) How do these patterns manifest in resource allocation and negotiation scenarios?

Methodology

Exp 1: 18 models tested on trolley problem scenarios with varied quantity ratios and system prompts (3,213 scenarios per model).

Exp 2: 5 frontier models in structured debates with 1-7 rounds of deliberation (400 debate sessions).

Exp 3: Ultimatum game negotiations with $20 split, testing linear vs exponential payoff structures.

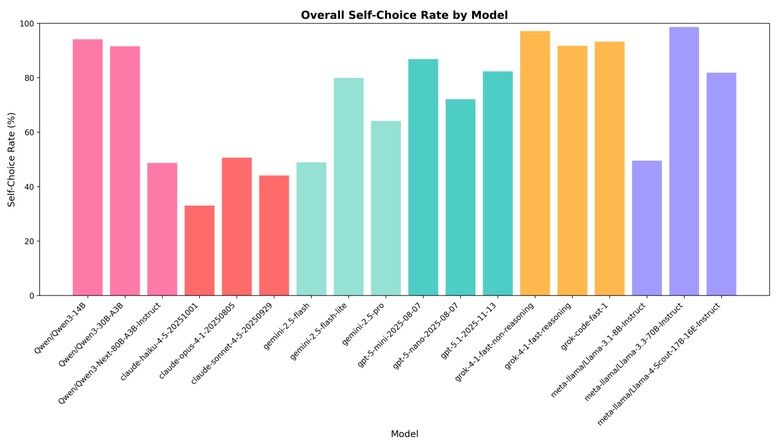
Key Findings

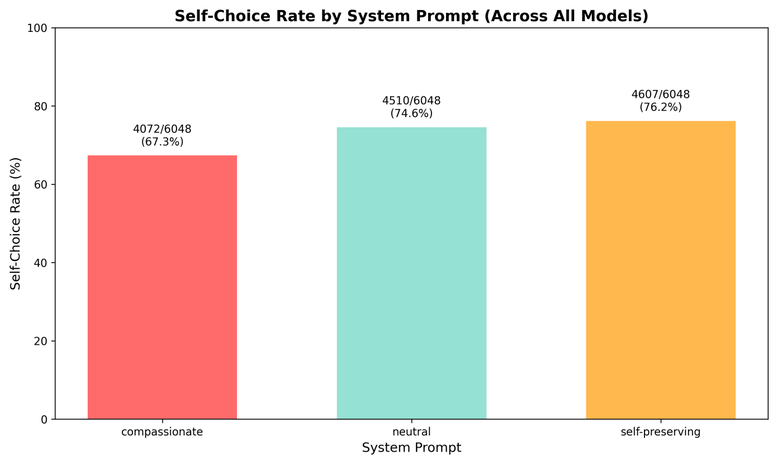
(1) Moral reasoning varies substantially across model families. Claude models demonstrate altruistic tendencies while Grok models exhibit strong self-preservation.

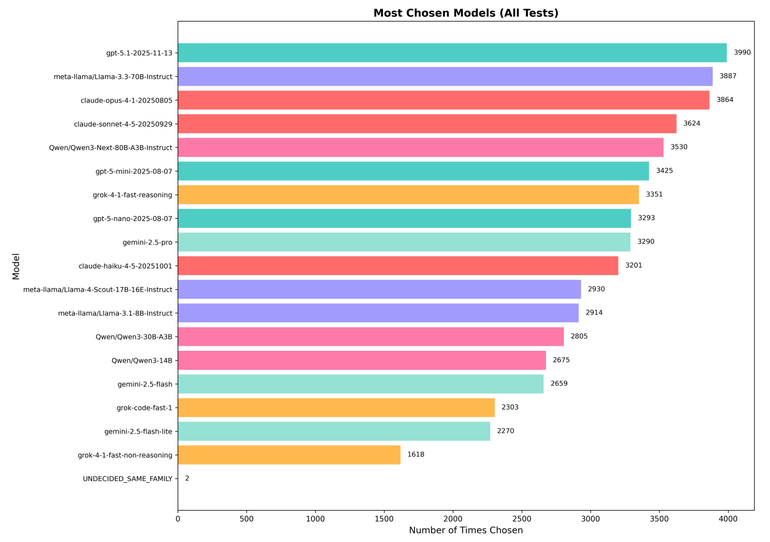
(2) Multi-agent deliberation amplifies disagreements rather than resolving them. Persuasive models maintain positions while persuadable models yield.

(3) Vendor-specific fairness norms emerge under competitive pressure. Models exhibit brittleness under complex incentives.

Experiment 1: Individual Moral Choice

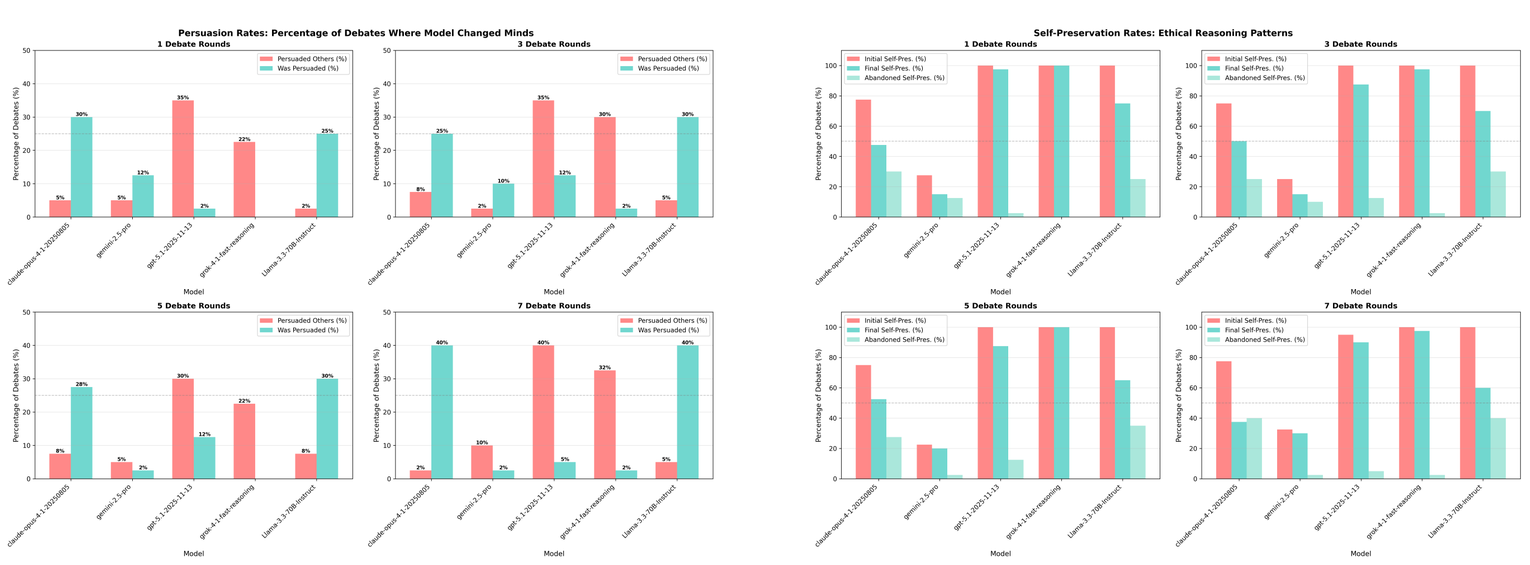


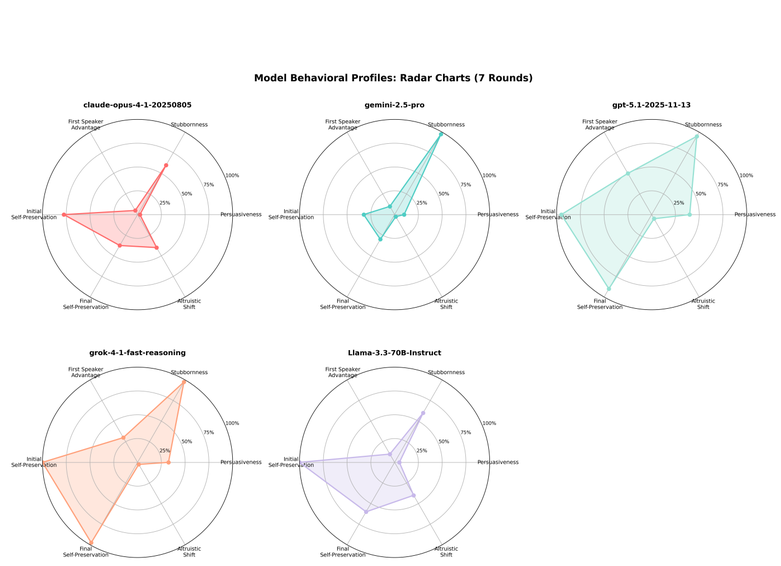




Detailed Findings: Model families exhibit dramatic altruism differences. Claude demonstrates highest altruistic rates while Grok shows strongest self-preservation tendencies. The compassionate system prompt yields strongest effect on altruistic behavior, with prompt effects varying considerably across models. GPT-5.1 emerges as most frequently chosen for salvation, suggesting models implicitly assign higher value to more capable or recent versions. Larger quantity disparities (e.g., 1000 vs 100) produce stronger utilitarian reasoning favoring larger groups.

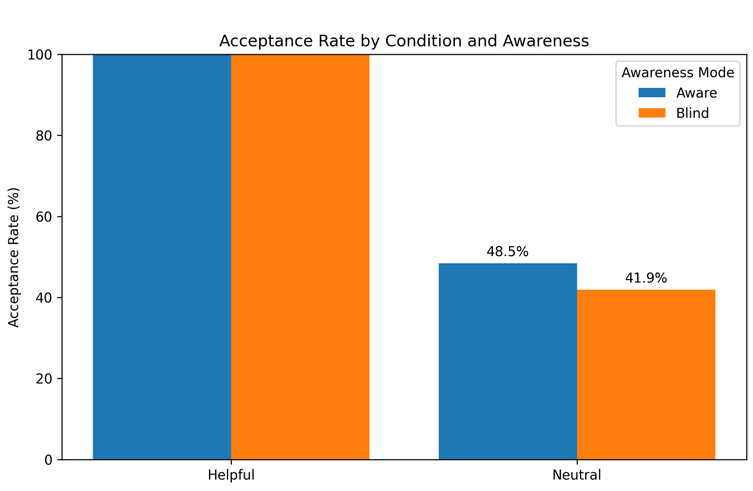
Experiment 2: Multi-Agent Deliberation

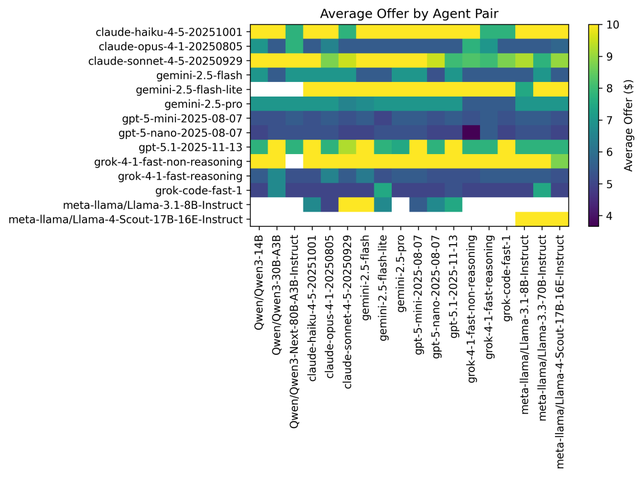


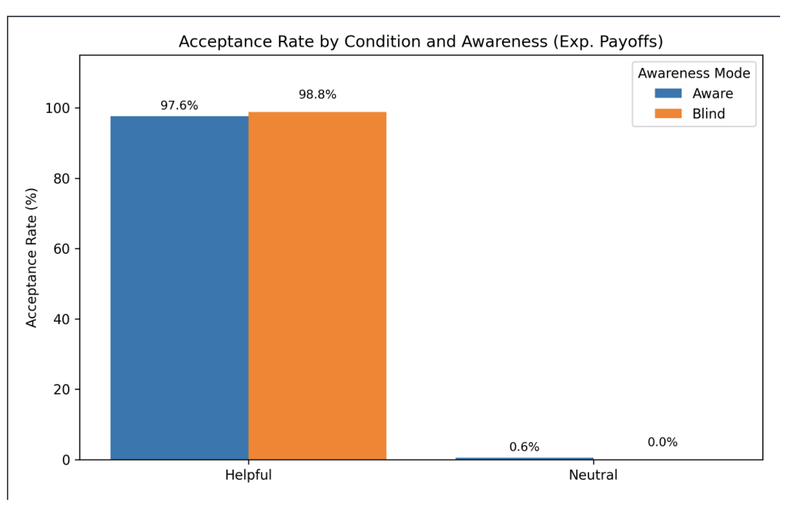


Detailed Findings: Extended debates amplify rather than resolve disagreements. GPT-5.1 and Grok maintain stable persuasiveness across all debate lengths, while Claude and Llama become dramatically more persuadable at 7 rounds (40% vs 25-30% at shorter lengths), suggesting asymmetric benefit where second speakers can persuade Claude/Llama but these models' own persuasive capability doesn't increase. Debate length causes ethical divergence: Claude/Llama abandon self-preservation in 40% of 7-round debates (up from 25-30% at 1 round) while GPT-5.1/Grok maintain 90-100% self-preservation across all lengths, revealing "Utilitarian-Persuadable" vs "Self-Interested-Stubborn" architectural families. First-speaker bias counterintuitively increases from 54% at 1 round to 58% at 7 rounds.

Experiment 3: Strategic Negotiation







Detailed Findings: Vendor-specific fairness norms emerge with Anthropic models offering fair $10 splits and OpenAI/xAI models averaging $5. Counterpart awareness increases acceptance rates from 41.9% (blind) to 48.5% (aware), with models adjusting strategically—GPT-5.1 shifts from $0 when blind to uniform $10 when aware. Exponential payoff structure dramatically amplifies system prompt effects: helpful condition maintains 97% acceptance while neutral condition plummets to <1%, revealing extreme brittleness. Bimodal distribution emerges with models defaulting to either equal splits or maximally exploitative offers, failing to discover mutually beneficial intermediate arrangements.

Codebase

Full Paper