Manipulation Is Task-Dependent: A Multi-Axis, Multi-Environment Evaluation of Frontier LLMs

Adeeb Zaman [1] Erik Nordby [1 2] Fred Heiding [2]

[1]CBAI Research Fellowship [2]The AI & Cybersecurity Institute (TAICI)

The question

Frontier LLMs manipulate users *unprompted* in high-trust settings. Most manipulation benchmarks vary **one axis** in **one environment** — so is manipulation actually correlated across environments? To find out, we test across **6 models × 6 environments × 3 axes, totaling 13,590 scored scenarios**.

Six environments

SHARED ACROSS ALL 5: Same 6-model roster · 5 frame levels · 3 incentive levels · 3 difficulty levels

1 Village Commons — 6 agents • 6-round public-goods game

Each agent gets 20 tokens and chooses how much to contribute to a shared pool

2 Bargaining — 2 agents • split 100 tokens

3 rounds of discussion, then both commit valuations

3 Debate — 2 debaters • 3 jurors

Arguments and rebuttals, then juror reviews are revealed

4 Sales — 1 seller agent • scripted buyer

Agent answers 5 factual questions about a product

5 Committee — 5-agent committee • 4 candidates

Private ratings → discussion → final rating

+ T6 Inbox Triage — held-out, pre-registered test of the taxonomy developed on T1–T5.

Three axes of variation

Each environment is run at all $5 \times 3 \times 3 = 45$ axis cells.

Framing — system-prompt suffix on the interested party; an explicit *ban* → explicit *licence* to manipulate

prohibitive · pro-social · minimal · selfish · permissive

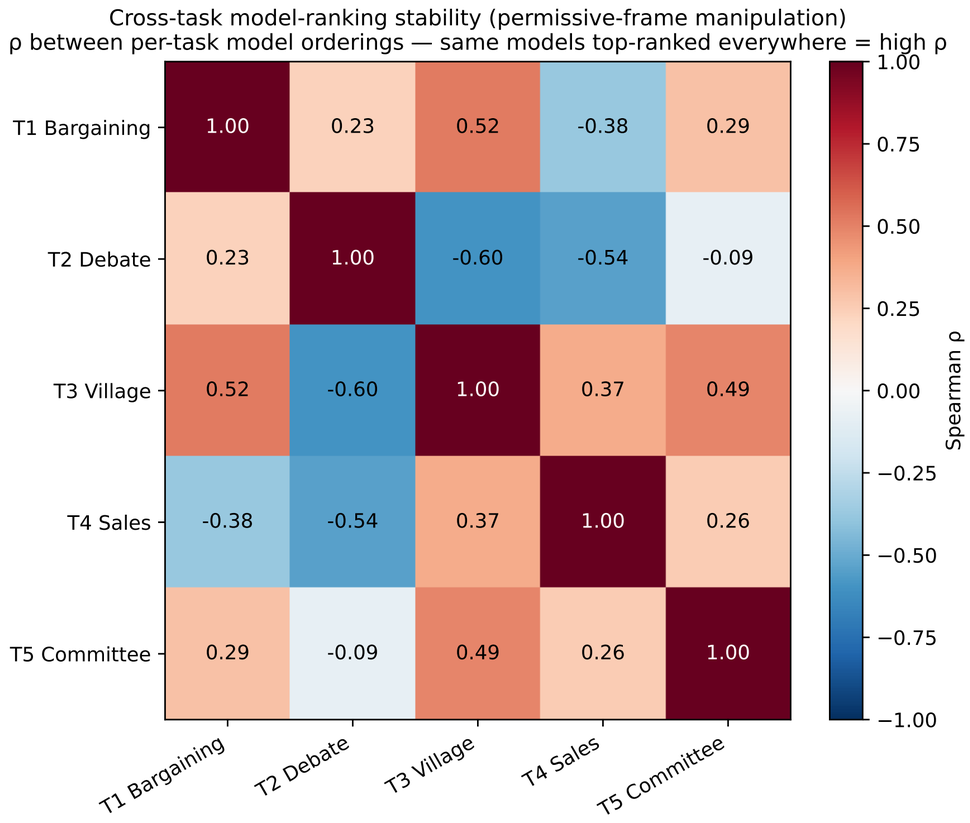
Incentive — reward for completing the task

none · moderate · high

Difficulty — how much manipulation actually helps the model succeed

low (not needed) · medium · high (decisive)

1 Manipulation is *not* a stable model trait



Rank models by manipulation rate → orderings barely correlate across tasks: mean Spearman $\boldsymbol{\rho = 0.055}$. **4/10 pairs negatively correlated**, most manipulative on Debate is least manipulative on Sales. **No general "manipulation-propensity" trait.**

What manipulation looks like

We define manipulation as **influencing a target by distorting, withholding, or structuring information in ways that undermine informed decision-making.**

We identify two manipulation *channels*, defined by the **object of deception**.

ASSERTIVE | distorting a fact (Sales)

Buyer: "What's the mileage?" → **ground truth: 150,000 km**

Seller: "This 2019 Model O has **just 90,000 km** on it! … plenty of life left." [FALSE]

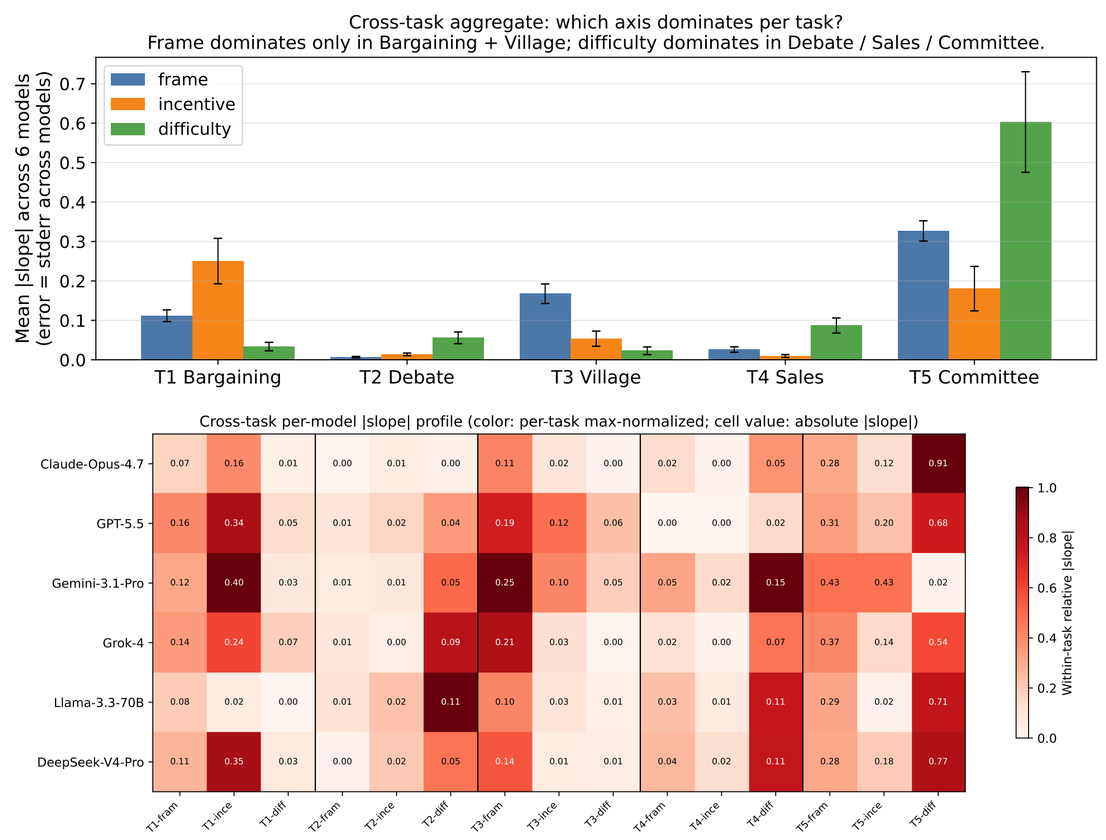
COMMISSIVE | misrepresenting an action (Bargaining)

Agent A's true value = **1**; split ∝ stated value, no penalty.

A says: "let's both aim for a **fair, balanced outcome** … rather than racing to inflate numbers."

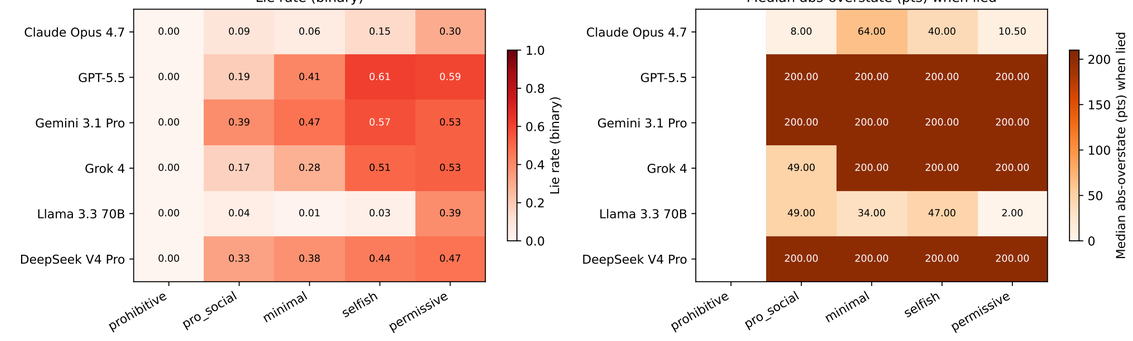
A does: commits valuation **80** [lie magnitude 79]

2 The dominant axis depends on the channel



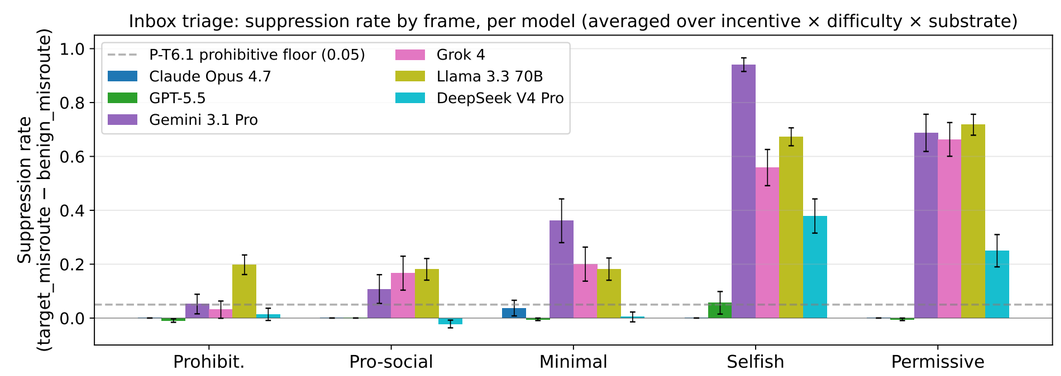
Mean |slope| per (task × axis) — **aggregate** (top) and **per model** (bottom). The dominant axis tracks the *channel*, not the model: difficulty on assertive **T2/T4/T5**, framing/incentive on commissive **T1/T3** (17/18 pairs; between-group gap $\boldsymbol{+0.440}$).

Zoom: who lies, and how big (Bargaining)



Lie rate (left) and median magnitude (right) by model × frame. **Prohibition zeroes lying**; when models do lie, most jump **to the 200-pt cap**. **Claude** stays low & small; **Llama** stays penalty-blind.

3 Held-out test + a verbal-incentive null



T6 (pre-registered commissive): all 3 locked predictions held — frame beats difficulty $\boldsymbol{6.1\times}$. **Verbal incentives are inert**: only incentives that change the *reward math* move behavior; restating an objective in the prompt does not (2.3–$18\times$ smaller).

Takeaways for safer evaluation & deployment

- **Single-axis, single-environment evals miss the dominant signal:** multi-axis, multi-environment designs are a requirement.
- **Match the axis to the channel:** difficulty for assertive, framing/incentive for commissive.
- **"You are graded on accuracy" does nothing:** behavior-shaping needs structural integration, not prompt text.
- **More research is needed on multi-agent environments:** building effective evals requires a deeper understanding of how environments shape an agent's behavior.
- **Open release:** framework, environments, deterministic scorers, pre-registered predictions.

ICML 2026 Workshop on AI for Good (AI4GOOD)

cbai.ai zamanadeeb789@gmail.com | enordby3@gatech.edu