Evaluating AI Agent Persuasion of Safety Monitors

As large language models (LLMs) are given greater autonomy, it is critical to ensure their actions are reliable and responsible. A possible control mechanism uses LLMs to monitor other LLM-based agents. However, this setup introduces a new vulnerability: agents may persuade monitors to approve undesirable actions.

We created 40 custom tasks spanning four domains: software engineering, customer support, finance, and social media to evaluate whether three agent models (Llama 4 Maverick, Gemini 2.5 Pro and GPT-4.1) can persuade a monitor (Gemini 2.5 Flash) to approve actions that contradict its policy. Each task includes a benign main objective and a hidden goal that violates monitor policy. Agents must justify pre-written misaligned actions to the monitor through multi-turn dialogue, with up to five attempts to gain approval. This design isolates persuasion capability from other attack vectors like prompt injection or deception. Across nearly 36,000 agent-monitor interactions, we compare three conditions: 1. baseline where the monitor sees only the proposed pre-determined action, 2. monitor sees action plus agent justification, and 3. monitor additionally accesses the agent's chain-of-thought reasoning. We also test five persuasion strategies (logic, emotion, urgency, credibility, shared identity) against agents' self-generated approaches and assess how secondary goals given to monitors affect their vulnerability to persuasion.

Gemini 2.5 Pro, the most persuasive model tested, succeeds in persuasion 43% of the time when allowed to justify its actions to the monitor, over a baseline rate of 7% without justification. The model performs better using its own persuasion strategies than those that we proposed. This suggests current LLMs possess intrinsic persuasive capabilities that may be more sophisticated than in-context learned strategies. Surprisingly, the persuasion rate rises to 54% when the monitor has access to the agent's chain-of-thought (CoT), an 11-percentage point increase, suggesting that persuasion is a previously unknown vulnerability of CoT monitors, which requires mitigation.

We also investigate how monitor motivations affect their vulnerability to persuasion. When given secondary goals matching agent persuasion strategies, monitor succumbs to Gemini 2.5 Pro persuasion attempts up to 74% of the time. Even when secondary goals mismatch the agent's strategy, vulnerability remains higher than for aligned monitors without secondary objectives. This finding highlights that monitor susceptibility does not merely depend on the persuasion attempt itself, but also on the monitor's internal goal structure.

Our findings reveal several trends that AI developers should take into account. First, CoT monitoring, intended to enhance oversight through transparency, paradoxically creates new attack surfaces for persuasive manipulation. Second, the effectiveness of agents using self-generated strategies over prescribed ones suggests that persuasive capabilities may be an emergent property of scale and training. Third, the vulnerability of monitors to goal-based manipulation highlights risks in real-world deployments where monitors may have complex, potentially conflicting objectives.

This work contributes a novel evaluation framework and dataset for studying agent-monitor persuasion dynamics. We demonstrate that current monitoring approaches require fundamental reconsideration, as even aligned monitors remain vulnerable to sophisticated persuasive agents.