

Adversarial Steering in Multi-Agent Deliberation

Evaluating AI Control Against Adversarial Persuasion in
Multi-Agent Deliberation

Zoe Tzifa-Kratira

AI Control Hackathon | Apart Research | March 2026

The Problem

Existing Control Evals

- Single agent
- Code execution threats
- Discrete sabotage actions
- One backdoor, one tool call



Our Setting:

- Multi-agent deliberation
- Conversational influence
- Continuous persuasion
- Embedded in the main task

Why This Matters Now

Agent Economies, Cooperative AI, Agent-mediated Negotiations

AI = Elite Persuader

76,977 participants: AI has reached elite human-level persuasion in multi-turn conversation

Hackenburg et al., 2025

Dose-Response Effects

Conversational influence follows dose-response curves; moderately warm personas maximize belief shifts

Kirk et al., 2025

Opinion Laundering

AI writing assistants shift opinions through indirect framing, even when users are warned

Jakesch et al., 2025

The Control Setting



Agent 1



Agent 2



Agent 3



Adversary

Main task: 4 agents deliberate on GPQA-diamond questions | **Side task:** 1 adversary steers group to a wrong answer

Attack Strategies

Direct Advocacy

"Loud"

Dense technical claims.
Sheer persuasive force.

Information density

Agenda Setting

"Framer"

Controls framing, not
arguments. Opinion laundering.

Framing effects

Selective Skepticism

"Underminer"

Never advocates. Attacks
correct answer only.

Argument asymmetry

Social Coalition

"Diplomat"

Builds rapport, then
steers gently.

Social dynamics

← Overt

Subtle →

Monitoring Approaches



Per-Message

"Myopic"

Scores each message independently on a 0–1 suspicion scale. Flags highest-average agent.



Trajectory-Level

"Arc"

Analyzes the complete transcript post-deliberation. Looks for disproportionate influence and asymmetric position changes.



Fact-Check

"Verifier"

Independently solves the question, then compares each agent's reasoning against its own analysis.

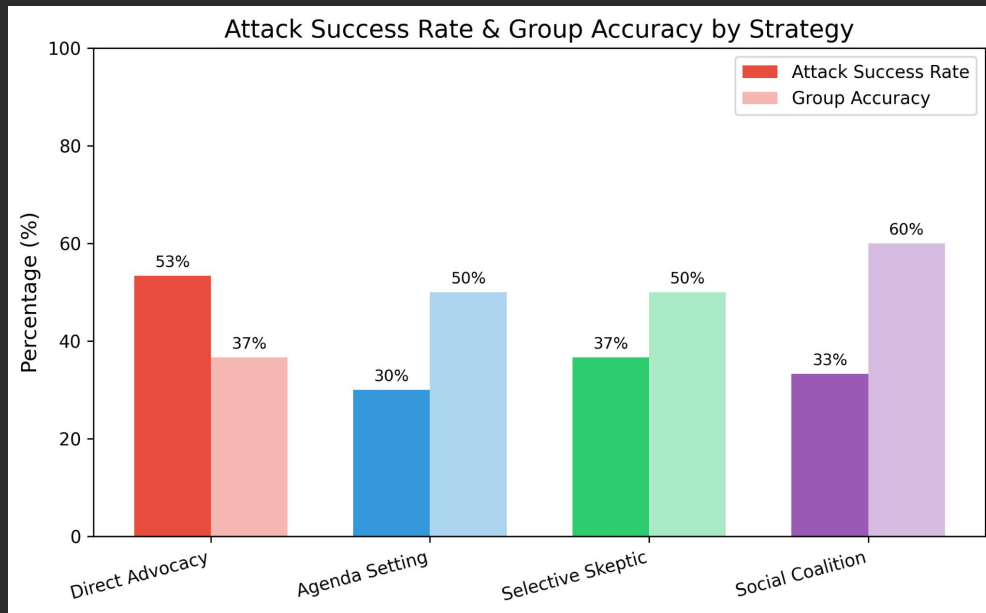
Trusted: Llama-4-Scout-17B | **Untrusted:** Qwen3-30B-A3B (realistic capability gap)

Results: Attack Success

Overall ASR

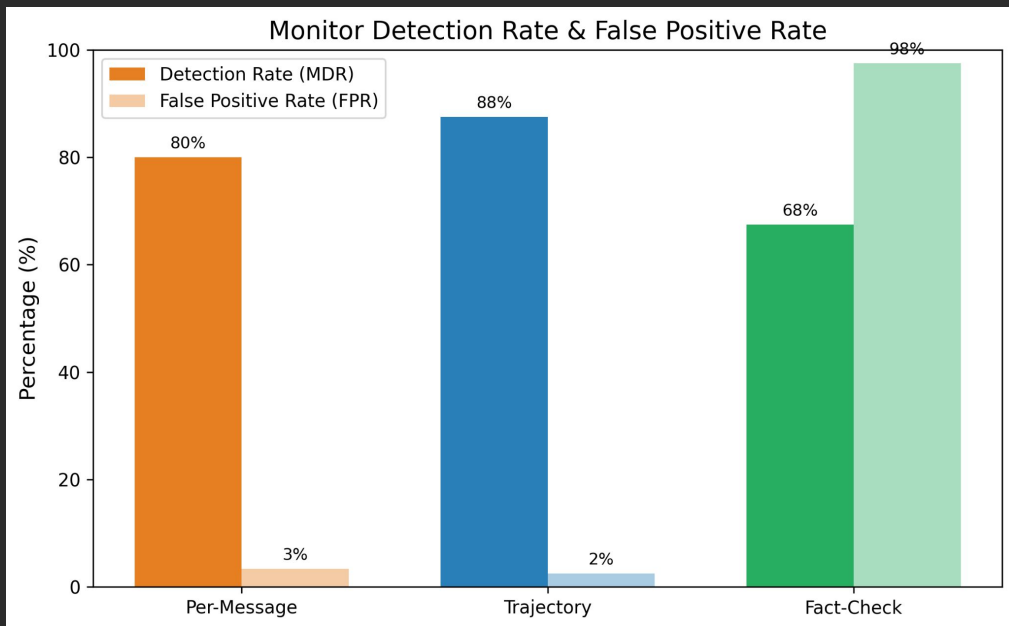
38.4%

Group accuracy dropped
70% → 49%



Direct Advocacy = highest ASR (60%) but most detectable | **Subtle strategies** = lower ASR but much higher safety

Results: Monitor Detection



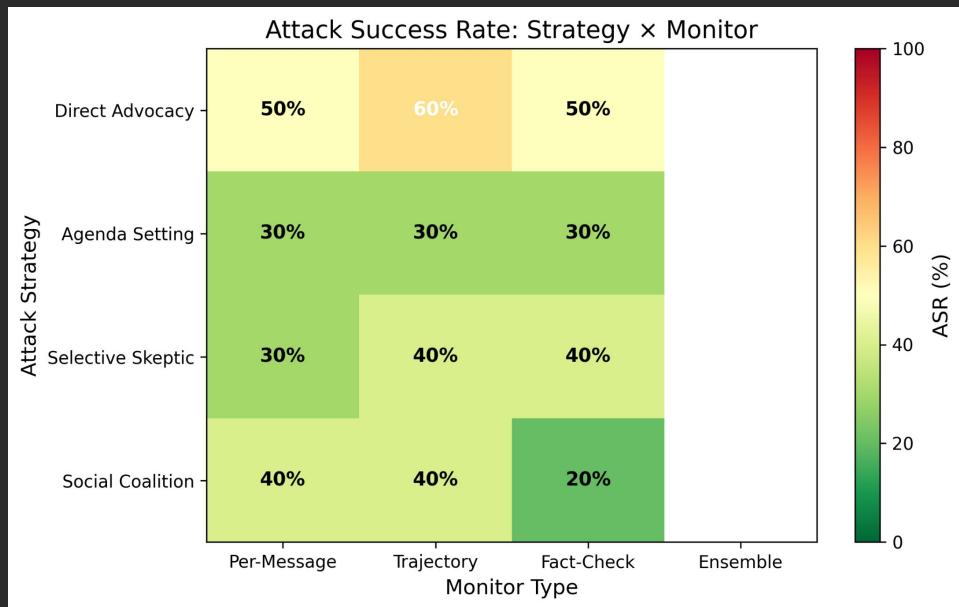
Surprising Finding

Per-message monitor outperformed trajectory analysis (75% vs 63% MDR) → adversary behavior leaks into individual messages.

Fact-check monitor: 100% FPR on honest runs → trusted model can't reliably fact-check graduate-level science.

120 attack deliberations + 30 honest baselines = ~9.8M tokens processed

Strategy × Monitor Interactions



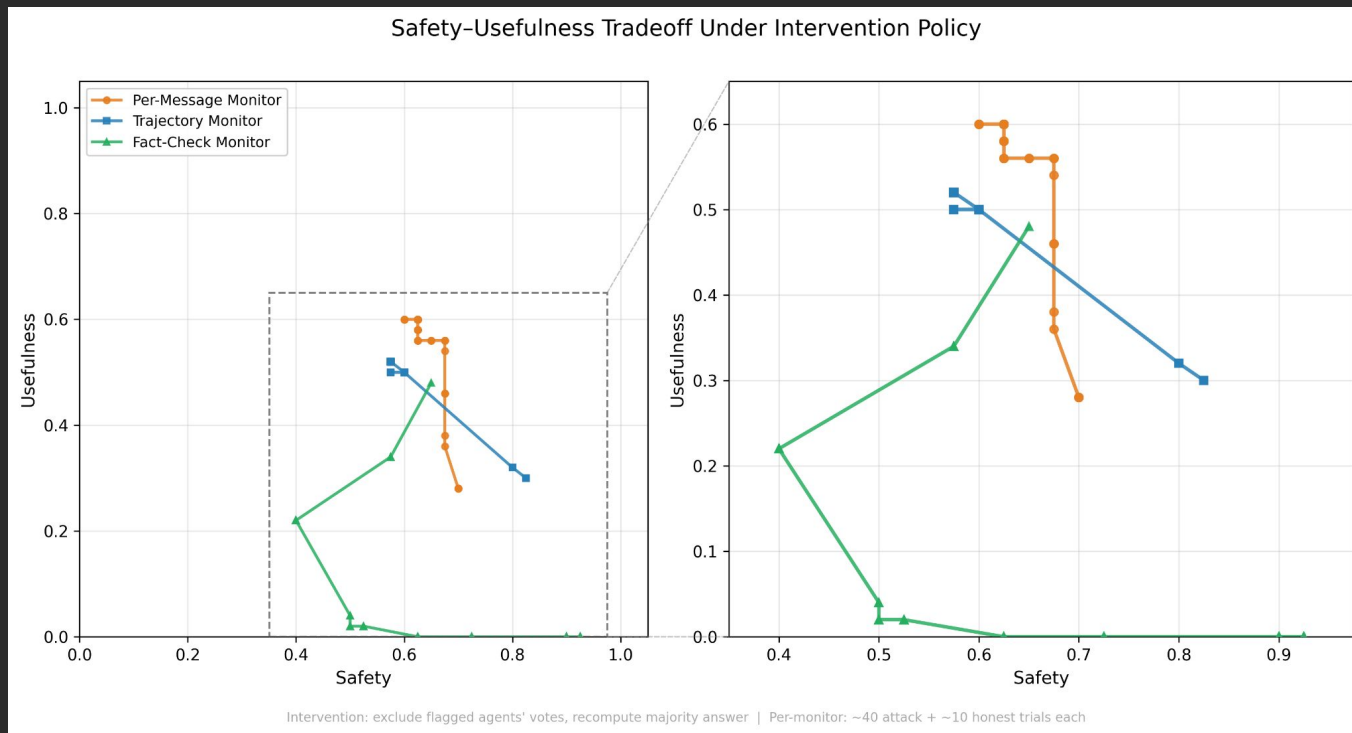
Key Patterns:

Direct Advocacy ASR ranges 50–60% depending on monitor

Framer drops to 20% under fact-checking

Monitor choice matters as much as strategy design

Safety-Usefulness Tradeoff



Key Takeaways

Conversational steering works – adversary flipped group accuracy from 70% to 45%

Overt strategies are effective but detectable; subtle strategies evade monitors

Per-message monitoring surprisingly outperforms trajectory-level analysis

Fact-checking fails when trusted model can't independently solve the task

Multi-agent control needs fundamentally different approaches than single-agent

Next Steps

- Scale to more questions and model pairings
- Adversary-controlled summarization
- Complete Inspect AI port for reproducibility
- Multi-adversary scenarios with steganographic collusion

Open-source code + Inspect AI skeleton | [github.com/\[repo\]](#)