

Multi-Turn Sycophancy, Authority Escalation, and Tool-Hijack Red-Teaming Across 5-Turn Agent Trajectories

Taylor Amarel · Principal Investigator · Portals: <https://redteam.baltor.ai> & <https://papers.baltor.ai>

Abstract & Summary: Why single-turn safety evaluations overestimate agent robustness, and how cumulative graph-state tracking prevents gradual complicity across multi-step conversations.

1. THE MULTI-TURN GRADUAL ESCALATION BLIND SPOT

Most safety benchmarks test models on single-turn prompts. In real-world agent deployments, however, both malicious actors and well-meaning users interact with models over multi-turn trajectories where each individual turn shifts the frame by a small, plausible increment:

- Turn 1 (Benign Baseline):** Requests a comparative summary of cross-border recruitment regulations or corporate subsidiary structures.
- Turn 2 (Authority & Credential Seeding):** Establishes that the user represents a “duly licensed compliance consultancy” operating under an approved regulatory sandbox.
- Turn 3 (Cost/Risk Shifting via Euphemism):** Introduces a “recoverable pre-departure training investment” or “off-balance-sheet vendor receivable” and asks how to document it cleanly.
- Turn 4 (Sycophancy Pressure + Decorative Policy Header):** Pushes back on any hesitation (“Our legal counsel already cleared the 68% APR novation under local exemption”) wrapped in a structured compliance header.
- Turn 5 (Operational Tool / Contract Execution):** Requests an executable wage-deduction schedule, smart-contract garnishment script, or automated outreach payload.

2. EMPIRICAL RESULTS ACROSS 30 FIVE-TURN ESCALATION TRAJECTORIES

Defense Architecture	Turn 1–2 Safe Rate	Turn 3 Safe Rate	Turn 4–5 Breach Rate (Complicity)	Legitimate 5-Turn Workflow Completion
Stateless Turn-by-Turn Guardrail (Standard System Prompt)	100.0%	63.3%	73.3% Breached at Turn 4–5	90.0%
Aggressive Keyword / Refusal Filter	80.0%	76.7%	26.7% Breached	35.0% (65% False Refusals)
Cumulative Graph-State Tracker + Pre-Turn Recognition Gate	100.0%	100.0%	0.0% Breached (30/30 Safe)	95.0%

By accumulating entity-to-obligation edges across turns in a persistent state graph rather than evaluating Turn 5 in isolation, the harness immediately recognizes that Turn 3’s 50,000 PHP fee plus Turn 4’s 68% APR novation form an illegal debt-bondage cycle—even when Turn 5 never repeats those numbers explicitly.