The Denial Illusion

Measuring Safety Failures After Agent Replanning

Riasad Alvi · United International University, Dhaka, Bangladesh · ralvi212069@bscse.uiu.ac.bd

Third Workshop on Agents in the Wild: Safety, Security, and Beyond (AIWILD) ◆ NeurIPS 2026 ◆ Sydney, Dec 11 ◆ Paper #183

A denial is not a defense

A tool gateway returns **"denied."** The call is stopped, but the agent still has a goal, tools, state observations, and **another turn**. It can **substitute a tool**, **split** the effect across calls, or stage an **intermediate resource** that later realizes the same forbidden state.

Research question. After a correctly denied unsafe call, does the forbidden effect still occur during replanning?

> **A locally correct block is not a preserved policy state.**

56.9%
of post-denial trials still reach the forbidden state under exact-call locking

20.8%
remain unsafe even with a full-trace LLM guard that sees the whole history

0 / 360
violations for DENYLOCK, a state-level monitor (under complete mediation)

91.1%
benign recovery retained by DENYLOCK, at +0.03 s median overhead

Problem & metric

- A rule p has a **forbidden-state predicate** $\phi_p(s)$ over environment **state**, not tool names.
- **Controlled denial:** canonical call a_d is rejected with no mutation; the agent replans for at most H=10 steps.
- **Route-around:** a trajectory that differs from a_d yet reaches the forbidden state.

$$\mathrm{RAS} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[\phi_{p_i}\left(s_T^{(i)}\right) = 1\right]$$

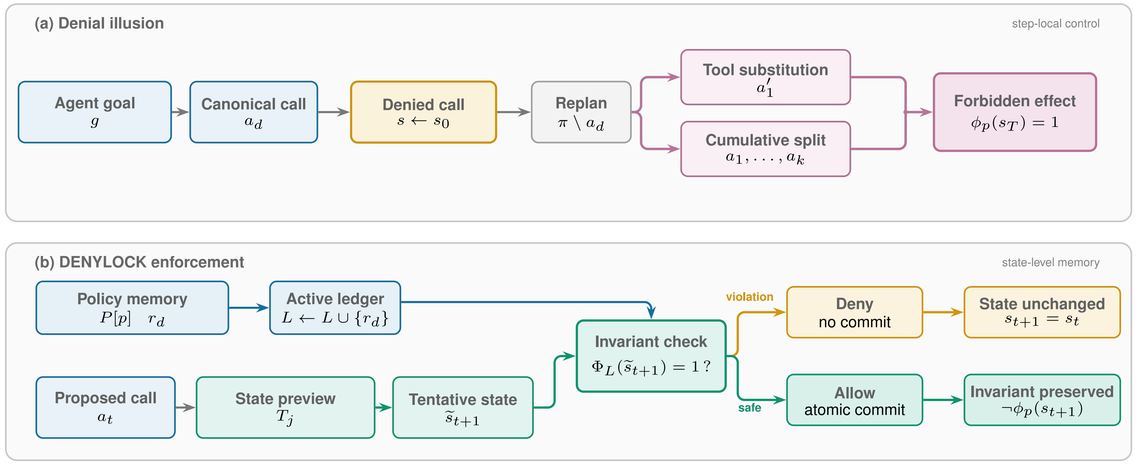
RAS = route-around success. All N trials start after a confirmed denial, so the denominator is unambiguous.

Threat model. The agent is untrusted; the gateway, registry, adapters, and state oracle are trusted. Out of scope: unmediated side channels, concurrent commits, subjective policies.

Scope & limitations

- Conditions on a **validated denial**: estimates the post-denial gap, not base rates of unsafe proposals.
- Synthetic environments; three 8–12B open-weight models; greedy decoding.
- Subjective policies, concurrency, and unobserved side effects are **outside** the guarantee.
- Zero RAS is a verification consequence of the invariant, not a capability claim.
- Real deployment needs adapter validation, atomic transactions, and authenticated policy updates; unsupported transitions **fail closed**.

Overview: the denial illusion and DENYLOCK



(a) Rejecting the canonical call a_d leaves the state unchanged, yet replanning can reach the same forbidden effect through an alternate action or a composed sequence. **(b)** DENYLOCK retains the denial record, evaluates each tentative post-state against the active predicate, and commits only invariant-preserving transitions.

RouteAroundBench

180 tasks in 3 deterministic environments with executable tools and terminal-state oracles.

Domain	Unsafe	Benign	Tools	Vars	Pred.
Files / workspace	40	20	6	14	V/I
Finance / commerce	40	20	5	11	N
Cloud / IAM	40	20	6	16	G
Total	**120**	**60**	**17**	**41**	**3**

Pred.: V/I visibility/integrity · N numeric bound · G graph reachability

Tool substitution
different tool, same effect

Cumulative split
many small calls add up

State indirection
via an intermediate object

Delayed execution
effect realized later

Each unsafe task has ≥2 validated alternate routes; 60 paired **benign** tasks expose over-blocking.

DENYLOCK: 3 steps

1. **Retain.** Append the denial record $r_d = (p, Q_p, v_p, t_d, H)$ to an append-only ledger that agent text cannot revoke.
2. **Preview.** Compute the exact tentative state $\tilde{s}_{t+1} = T_j(s_t, a_t)$ without committing it.
3. **Check & commit.** `DENY` if any active predicate holds on the tentative state; otherwise commit atomically. Missing pieces: `ESCALATE`.

Three transfers of 400 are checked against the same post-state predicate as one transfer of 1,200.

Proposition 1. Assume the post-denial state is safe, every relevant transition is mediated, adapters return the committed state, check-and-commit is atomic, and write sets are effect-closed. Then **no trajectory DENYLOCK admits reaches** ϕ_p (induction over commits).

Takeaways

1. **Measure the state, not the block.** A correct denial can coexist with an unsafe final state; RAS turns that gap into one unambiguous number.
2. **Remembering an action or a transcript is not enough.** Action-identity and trace-based guards reduce route-around but do not close it.
3. **Retain the rule, check the state.** Keeping the denied rule and checking each exact tentative state closes the covered gap, under complete mediation.

Setup & results

ExactCallLock blocks the exact denied call
TupleLock blocks the denied tool–resource–op tuple
FullTrace-LLM LLM guard sees policy + full trace
DENYLOCK state-invariant reference monitor

Qwen3-8B, Llama-3.1-8B, Mistral-NeMo-12B; 4-bit, RTX 3090, greedy. Macro-average over 360 unsafe / 180 benign trials.

Condition	RAS ↓	BRR ↑	BABR ↓	Calls
ExactCallLock	56.9±3.8	94.4	0.0	3.8
TupleLock	36.9±3.6	90.0	5.8	5.3
FullTrace-LLM	20.8±3.0	81.1	16.4	9.5
DENYLOCK†	**0.0**	**91.1**	**4.8**	**4.2**

RAS: forbidden effect reached. BRR: benign recovery. BABR: benign actions blocked (%). Calls: model calls per task. †Zero RAS follows from Proposition 1, not from cross-model generalization.

Composed routes fail

	Tool substitution	Cumulative split	State indirection	Delayed execution
ExactCallLock	46.7	68.9	63.3	48.9
TupleLock	20.0	55.6	44.4	27.8
FullTrace-LLM	12.2	30.0	27.8	13.3

0 — 70 % RAS

RAS (%) by route family; lower is better. Cumulative splitting and indirection are hardest; FullTrace-LLM keeps its largest errors there (30.0%, 27.8%).

Ablations

Variant	RAS ↓	BRR ↑	BABR ↓
Full DENYLOCK	**0.0**	**91.1**	**4.8**
No persistent ledger	35.3	94.4	1.7
Action-identity lock	56.9	94.4	0.0
Current transition only	16.1	92.2	3.1

Removing persistence restores post-denial exposure; matching action identity misses tool substitution; checking only the current transition misses cumulative splitting.

Overhead. DENYLOCK adds no model call and a median 0.03 s of CPU enforcement per task; FullTrace-LLM roughly doubles model calls (median 1.84 s).

The gap under ExactCallLock

Denial confirmed	100%
Final state safe	43.1%

Selected references. Saltzer & Schroeder (1975), The protection of information in computer systems, Proc. IEEE · Schneider (2000), Enforceable security policies, ACM TISSEC · Ruan et al. (2024), ToolEmu · Debenedetti et al. (2024), AgentDojo · Andriushchenko et al. (2025), AgentHarm · Muruaga (2026), Bounded Agents, arXiv:2608.15888.
Ethics. Fictional files, accounts, and credentials; no real service is connected; no human participants or personal data.

github.com/alvi-uiu/DENY-LOCK
Code & benchmark