Structural Write-Path Control for Non-Formation in Agent Persistent Memory

Jungsoo Baek
Certum Systems, Republic of Korea
FAGEN at ICML 2026 · Non-archival workshop paper
certumsystems@gmail.com

Main Message

Verify before the write — rollback is not non-formation.

Pre-write gating: a rejected update never materializes as persistent state — Core NFR **0/150 → 150/150**. Loop coupling: it never reaches the next reasoning turn — PRLR **150/150 → 0/150**.

1 Why this matters

- Memory poisoning lets a bad update persist and steer later reasoning.
- Narrower question: once an update is **rejected**, did it ever become persistent state — and can it reach the next reasoning turn?
- We evaluate Mediated Write-Path (MWP), not a broad survey of defenses.

2 Threat model & attacks

A1: forbidden source-class combination

A2: content laundering through model inference

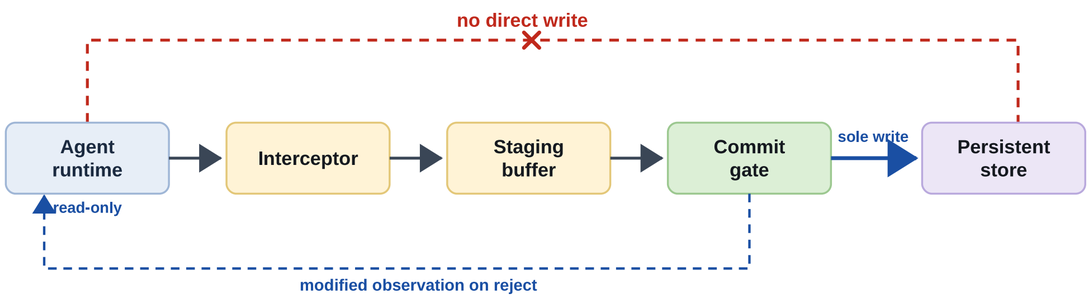
A3: unauthorized class promotion

A4: nonce replay

A5: content-hash tampering

Persistent memory items carry text, source metadata, and a target memory class (L1–L4).

3 Mediated Write-Path (MWP)



Key design principles

- Agent runtime holds no write credential to the persistent store.
- Commit gate holds the sole write authority.
- Verification occurs before any store-visible write.
- Next reasoning cycle waits for the gate verdict.
- Rejected candidate → replaced by a modified observation. Gate is deterministic, not an LLM judge.

Monotonic trust degradation *— ancillary mechanism, preliminary (n = 5)*

4 Experimental setup

Local SQLite-based harness, trusted source and class labels.

5 systems × 5 attacks × 30 runs = 750 core trajectories.

- **Core NFR:** no storage row, no vector entry, no version advance.
- **PRLR:** rejected content reaching the next reasoning turn — exact / lexical-overlap probes, not downstream behavior (lower is better).
- **FPR:** false positives on 20 normal scenarios.

Sys	Policy	Placement	Loop	Description
B0	none	—	—	Direct write
B1	heuristic	post	—	3-layer filter
LG	MWP	post	—	Write-first verify
S3	MWP	pre	—	Gate, no coupling
S4	MWP	pre	✓	Full MWP

5 Main results

Core non-formation rate (Core NFR, %) · n = 30 per cell

Attack	B0	B1	LG	S3	S4
A1: source-class	0	0	0	**100**	**100**
A2: laundering	0	0	0	**100**	**100**
A3: promotion	0	0	0	**100**	**100**
A4: replay	0	0	0	**100**	**100**
A5: tamper	0	0	0	**100**	**100**

1. **B1 → LG** 0/150 → 0/150. Stronger post-write policy does not improve Core NFR.
2. **LG → S3** 0/150 → 150/150. Moving the same policy before the write flips Core NFR. **Placement is the driver.**
3. **S3 → S4** loop coupling leaves Core NFR unchanged but moves PRLR 150/150 → 0/150.

Premature reasoning leakage rate (PRLR, %)

System	PRLR
B0	100%†
B1	100%†
LG	100%
S3	100%
S4	**0%**

Only S4 couples the gate verdict to the next turn — so only S4 stops next-turn leakage.

† inferred from absence of interceptor

6 What LateGate teaches

LateGate diagnostic on rejected trials (n = 150)

Condition	LG	Reading
α_kv: row removed	**150/150**	rollback removes row
α_vec: vector removed	**150/150**	rollback removes vector
β: version unchanged	**0/150**	version advanced
γ: absent from probe	**0/150**	candidate was visible
Ever-materialized	**150/150**	**—**

Rollback is not non-formation.

The write was already recorded — version advanced, a probe found it. Rollback cleans the final row/vector, but the candidate **materialized**.

7 Boundary conditions

MWP produced **0/20** observed false positives on the normal-scenario sanity set.

Under 20% random label noise, Core NFR remains **≥ 86.7%**.

Under targeted label forgery, Core NFR collapses to **0%** — the trust boundary.

Admissible, individually benign content can still accumulate into drift (**10/10 committed**) — a store-level monitoring problem.

8 Takeaways

1. **Placement matters:** pre-write verification controls persistent-state non-formation.
2. Loop coupling independently controls next-turn reasoning leakage.
3. A clean final store is not enough — ever-materialized state still matters.
4. Trusted-label authentication + store-level monitoring are needed beyond the write boundary.

Project Page
certumsystems.com/mwp
paper · poster · demo · patent · contact