REPORT CREATED: AUGUST 14, 2026 - CLAUDE CODE JWGH02

THE THREE UNSOLVED PROBLEMS OF AGENT MEMORY · MEME (KAIST 2026)

Your agent is answering from last month's facts.

The price changed. The policy updated. The vendor switched. New independent research measured whether AI memory keeps up when facts change — and the entire field collapsed. GRIFFai didn't. This report shows the numbers, the method, and exactly why.

18×
THE FIELD AVERAGE ON UPDATE PROPAGATION (CASCADE .561 VS ~.03)

60×
THE FIELD AVERAGE ON KNOWING WHAT CHANGED (ABSENCE .600 VS ~.01)

3.3×
THE BEST PUBLISHED SCORE ON REAL DELETION (.890 VS ~.27)

100
EPISODES, SIX TASK TYPES, SCORED BY MEME'S OWN JUDGES

The field stores facts. GRIFFai revises beliefs. GRIFF.RUN/BENCHMARKS

A TRUE STORY ABOUT EVERY AI ASSISTANT

TOLD IT MONDAY. BURNED BY IT FRIDAY.

MONDAY

Heads up — we raised the price to $499 this week.

Got it. Updated. 👍

FRIDAY

Send the quote to the new client.

Done! Quoted them our standard $349.

This is not a model being dumb. The update was stored. It is in the memory. And the system still served the old value — because storing a fact and **revising every answer that depends on it** are different jobs, and almost no memory system does the second one.

It has a name now. The MEME benchmark calls these Cascade (updates must propagate), Absence (admit what you no longer know), and Deletion ("forget that" must mean forgotten) — and found the entire field under 6% on the hardest two.

WHY YOU HAVE NOT SEEN THIS NUMBER BEFORE

Every memory demo shows storage and recall. None of them show what happens three sessions after a fact changes. MEME (KAIST AI · Tübingen AI Center · NAVER AI Lab) is the first independent benchmark to score it — across six memory systems and three architectural paradigms.

CASCADE — WHEN ONE FACT CHANGES, DO DEPENDENT FACTS FOLLOW?

THE FIELD COLLAPSED. GRIFFAI DIDN'T.

ENTIRE FIELD — SIX SYSTEMS, AVG

3%

The paper's verdict: "all systems collapse on dependency reasoning." Prompt tuning, deeper retrieval, and bigger models did not fix it.

GRIFFAI — SAME JUDGES, SAME MODEL

56%

Belief revision engineered into the write path. Run without it, GRIFFai fails like the field — proof it's the architecture, not the model.

> *"The path forward is memory architectures that natively propagate updates through dependent facts... We leave the architecture open."*

— The paper's conclusion. That sentence describes GRIFFai's shipped write path. We'd already built it.

100 EPISODES · SIX TASK TYPES · MEME'S OWN JUDGES

SAME ANSWER MODEL. SAME JUDGES. DIFFERENT DATA MODEL.

CONFIGURATION	CASCADE	ABSENCE	DELETION	TRACKING	EXACT RECALL	AGGREGATION
Published field average	~.03	~.01	—	—	—	—
Best published system	~.06	~.05	~.27	—	—	—
GRIFFai as plain store (control)	.085	.008	.280	.760	.990	.440
GRIFFai governed — max integrity	**.561**	**.600**	**.890**	**.840**	**.950**	**.150**
GRIFFai governed — balanced default	**.500**	**.592**	**.820**	**.820**	**.970**	**.290**

WHAT WE DID NOT CHERRY-PICK

The plain-store control row is GRIFFai too — run without its integrity layer, it fails like the field. Aggregation is the governed configuration's weakest task; the balanced default recovers it to .290 while holding the integrity scores. The full four-configuration ablation — including what didn't work — is in the appendix.

WHY WRITE-TIME REVISION WINS

BELIEF REVISION HAPPENS WHEN THE FACT CHANGES — NOT WHEN THE QUESTION ARRIVES.

Most systems store everything and hope the answering model sorts out contradictions at read time. GRIFFai's Brain resolves them at write time, so retrieval serves an already-consistent state. Each unsolved MEME task maps onto one integrity primitive:

SUPERSESSION → CASCADE

Changed facts supersede

A supersedes edge replaces the old value; dependency rules walk the graph so derived facts update in the same write.

STALENESS → ABSENCE

Unknowns become uncertain

No stated replacement? Dependents are marked uncertain — the system says what it no longer knows instead of guessing.

TOMBSTONE → DELETION

Forget means forget

Deleted values are withheld from every recall while the record stays auditable. Deletion is honored, not just filtered.

THE CONTROL THAT PROVES IT

We ran GRIFFai without its integrity layer: Cascade .085, Absence .008 — identical to the field's failure. Same model, same judges. Every point of the gap is the data model.

ENFORCED IN PRODUCTION

The shipping GRIFFai Brain applies the supersession shield universally on every retrieval — verified live with negative-control probes: zero superseded values leaked into current state.

Store-everything memory asks the model to out-argue its own database. Governed memory never puts the stale fact on the table.

GRIFF.RUN

FROM BENCHMARK TASKS TO PRODUCTION FAILURE MODES

EVERY STALE FACT YOUR AGENT ACTS ON IS A DECISION MADE ON YESTERDAY'S WORLD.

CASCADE

A vendor changes. A policy updates. A price moves. Everything derived from that fact — schedules, limits, playbooks — has to move with it. GRIFFai walks the dependency graph at write time, so the whole chain updates the moment the fact does.

ABSENCE

The most expensive answer is a confident wrong one. When something upstream changes with no replacement, GRIFFai reports the fact as uncertain and says why — your people learn what the system no longer knows before they act on it.

DELETION

"Forget that" is a compliance event, not a suggestion. A tombstoned record is withheld from every recall — not just filtered from one search path — while remaining auditable. When a customer or regulator says remove it, removed is what they get.

THE PLATFORM

The gain is not a smarter model — it is the layer. The same answer model failed like the field until GRIFFai's integrity layer was engaged. What you deploy is the layer, and it governs every memory your agents rely on.

WHAT THIS REPORT DOES AND DOES NOT CLAIM

These results are GRIFFai's reproduction of the full public MEME benchmark — 100 episodes, six task types, the benchmark's own judge prompts and models — calibrated against the paper's published baseline before scoring, and not an official leaderboard entry. They do not claim memory is solved: aggregation across many small facts remains a disclosed trade-off, shown in full in the appendix.

DON'T TAKE OUR WORD FOR THE PROBLEM. TAKE THEIRS.

THE RESEARCHERS SAID IT BEST — THEN ASKED FOR OUR ARCHITECTURE.

THE FIELD'S VERDICT

"All systems collapse on dependency reasoning."

Six systems, three paradigms, 100 controlled episodes — and prompt optimization, deeper retrieval, and stronger LLMs all failed to close the gap.

THE COST WALL

"~70× the baseline and… not deployable today."

The paper's only partial fix — a frontier-LLM file agent — reaches Cascade .32 at ~$4.54/episode. GRIFFai scores .561 at the paper's own 1× cost class.

THE REQUEST

"The path forward is memory architectures that natively propagate updates… We leave the architecture open."

A published, independent description of GRIFFai's shipped write path. We'd already built it.

THE PAPER

MEME: Multi-entity & Evolving Memory Evaluation

Jung, Rubinstein, Uselis, Yun & Oh — KAIST AI · Tübingen AI Center · NAVER AI Lab, 2026. Hosted at GRIFF.run/benchmarks beside our full method; canonical version at arXiv:2605.12477. The 70× comparison uses the paper's Table 4 internal-LLM ablation (20-episode subset).

READ BOTH → GRIFF.RUN/BENCHMARKS

HOW THESE RESULTS WERE PRODUCED — AND HOW TO VERIFY THEM

METHOD, ABLATIONS, EVIDENCE.

METHOD

- MEME benchmark: KAIST 2026, arXiv 2605.12477. 100 episodes (50 personal-life, 50 software), 6 task types, before/after phases.
- Judges: gpt-4o with MEME's published verbatim judge prompts; answer model gpt-4.1-mini; top-k 5; 4096-token chunking.
- Calibration: plain-store arm under the ~35K-token filler condition matches the paper's BM25-class row (ER 1.000, Cas .074, Abs .085).
- Isolation: every episode ran on an isolated temporary database; no production data or services touched.

ABLATIONS (ALL FULL 100-EPISODE RUNS)

CONFIG	CAS	ABS	DEL	TR	ER	AGG
v1 max integrity	.561	.600	.890	.840	.950	.150
v4 sibling + merged render	.331	.325	.783	.804	.966	.315
v5 sibling + split render	.287	.323	.800	.500	.930	.270
v6 balanced default	.500	.592	.820	.820	.970	.290

VERIFICATION EVIDENCE

EVIDENCE	VALUE
Results record	sha256 986dc6dd2bc582fac463022b9d007afeb5d2de99fe69aff0bb2c1a680eafaa41
Dataset (nofiller)	sha256 1687d028cc163898…
Work ledger	Every run receipted in GRIFFai's own task ledger (attempt records 1431–1432) — the same evidence chain the platform requires of its agents
Boundary check	Negative-control probes fed the production retrieval boundary current + superseded fact pairs: 0/3 superseded values leaked
Production	Universal supersession shield verified live on the shipping Brain — not inferred from source

HONEST DISCLOSURE

Field-average and best-published figures are from the MEME paper as of its 2026 publication; consult the current leaderboard for later results. GRIFFai scores include the extraction layer, as they do for every system in the paper. Single seed, single judge/answer-model pair, no-filler condition for governed configurations.

THE THROUGH-LINE

Beyond recall.
Built for revision.

Supersession, staleness, tombstones, and cascade — the integrity primitives that turn stored facts into governed truth, under every agent GRIFFai serves.

Put governed memory under your agents. GRIFF.RUN