Contents

Provenance-Tiered Conflict Resolution in Agent Memory

A component of agent context integrity: resolving conflicting stored facts by write-time trust

Yao Tsakpo Independent research · System studied: Inbin (inbin.dev) July 2026

Abstract

We use the term agent context integrity for the property that the information an autonomous agent communicates, stores, retrieves, and acts upon preserves enough provenance, attribution, verification, and conflict semantics to support reliable reasoning, and study one component of it: conflict resolution in long-lived memory. Autonomous agents increasingly answer from an accumulated store of facts rather than from a single retrieved passage, and that store, like any long-lived memory, comes to hold facts that conflict: a value the agent recorded last week, a newer guess it made this morning, a decision a human once handed it. When asked a question whose answer is contested, the agent must *resolve* the conflict, and the store gives it only the information it kept. We observe that mainstream agent-memory systems keep recency and relevance but discard the one signal that determines which conflicting fact to trust: where the fact came from. We propose attaching a discrete provenance tier to every fact at write time, *human-decided* > *source-verified* > *agent-asserted*, and resolving conflicts by tier rather than by recency. We give two closed-form models for the accuracy of a recency-resolving store versus a tier-resolving store as a function of conflict depth and conflict density, confirm them with an exact model-free simulation, and measure the effect on five instruction-tuned models across three architectures. The deterministic model is a scaffold, not the finding: under a deliberately adversarial construction it fixes the extremes (a recency-only store cannot resolve an adversarially-ordered conflict; a tier store resolves it by construction), setting up the question the empirical study answers, whether real models exhibit the same pull toward recency when the tier is withheld. That empirical result is directional across every model tested, though its magnitude depends on how well a given model engages the task, and

it is where the paper's weight rests. We situate this against trust-weighted retrieval, where reliability-based conflict resolution already exists for external web sources, and argue the contribution is its move into agent *memory* as a *discrete, write-time, unforgeable* tier: in particular the highest tier is minted only through an authenticated owner channel, so an agent cannot promote its own assertion to the tier that wins. We frame the broader property, verified at the source, attributed to its origin, resolvable to ground truth, as context integrity, and position this as one measurement of it.

---

1. Introduction

An agent is only as reliable as the context it acts on. A long-lived autonomous agent does four things with information, it communicates it (to other agents and services), stores it, retrieves it, and acts on it, and its reliability depends on the information surviving all four steps intact. In this work we use the term agent context integrity to refer to that property: that the information an agent communicates, stores, retrieves, and acts upon preserves enough provenance, attribution, verification, and conflict semantics to support reliable downstream reasoning. We do not claim this names a new field; we use it as the organizing frame for a program of measurements, each isolating one component.

This paper studies one component: conflict resolution in long-lived memory, once information has arrived and been stored, how should contradictions among stored facts be resolved. (A companion measurement, of the *communication* component, studies whether transmitting a reference rather than a payload lets a receiver reconstruct richer context from less transmitted information; §6 relates the two.) Our claim here is deliberately modest and empirical: we do not claim to introduce trust-based resolution, which exists (§2); we show that, *in current agent-memory designs*, recording provenance and resolving by it measurably improves correctness on conflicting facts, at negligible storage cost, and we characterize where it fails.

Most attention to context reliability has gone to *recall*, getting the right facts back into the context window, and the field has strong results there: hierarchical memory paging (Packer et al., 2023), vector retrieval, and summarization buffers all address remembering enough. But recall is not the whole of reliability. When the recalled facts *disagree*, and in any long-lived memory they eventually will, the agent faces a second problem: *resolution*, deciding which of several contradictory facts to believe.

The knowledge-conflict literature has established, repeatedly and across venues, that language models handle this badly. Xu et al. (2024) survey the area and organize it into context-memory, inter-context, and intra-memory conflicts. ClashEval (Wu et al., 2024) shows models override their own correct prior knowledge more than 60% of the time when handed incorrect retrieved content. WikiContradict (Hou et al., 2024) shows models fail to reflect contradictions faithfully, especially implicit ones. And the resolution the model *does* perform is governed by surface features, coherence, persuasiveness, position, rather than by any notion of which source deserves trust: Xie et al. (2024) find models are highly receptive to external evidence when it is coherent, and Liu et al. (2024) show accuracy depends heavily on *where* in the context a fact sits, a positional bias that in a memory store manifests as a recency bias.

This paper starts from a systems observation about that failure. In an agent-memory store, whether

a conflict resolves correctly is decided before the model ever sees the facts, by *what the store chose to keep*. A store that keeps only values and timestamps can offer the resolver nothing but recency. A store that also keeps each fact's provenance, the kind of origin it came from, can offer a principled order. Our claim is narrow and testable: recording provenance as a discrete tier at write time, and resolving conflicts by that tier, is both a small change to the store and a large change to correctness, precisely on the questions that conflict.

We make three contributions:

1. A write-time tier model (§3). Every fact carries a tier, *human-decided* > *source-verified* > *agent-asserted*, assigned by the channel it was written through, not by the writer's claim. We detail (§3.3) why the tier must be channel-derived: an agent that can label its own assertion "human" defeats the entire order, so the top tier is minted only through an authenticated owner channel.
2. Two closed-form accuracy models (§4), for conflict *depth* and conflict *density*, which fix the axis and extremes under adversarial seeding. We treat these as scaffolding: they are near-definitional once the seeding is fixed, and their role is to frame the empirical question, not to stand as the finding.
3. A multi-model measurement (§5), the paper's primary evidence, of whether real instruction-tuned models are pulled toward the recent wrong value when the tier is withheld. We test five models across three architectures, report the effect's direction and magnitude, and report honestly the models on which it is muted.

We are explicit about what is *not* new here (§6): trust-weighted conflict resolution already exists for retrieval, where RA-RAG (Hwang et al., 2025) resolves conflicts by a vote weighted by *estimated* per-source reliability, and Trust-Align (Song et al., 2025) aligns models to attribute and refuse. Our contribution is not trust-based resolution in general; it is its instantiation as a *discrete, write-time, unforgeable* tier inside *agent memory*, a setting those retrieval methods do not address, where one of the writers is the untrusted agent itself.

---

2. Background and related work

Knowledge conflict. The canonical reference is Xu et al., *Knowledge Conflicts for LLMs: A Survey* (EMNLP 2024), which taxonomizes context-memory, inter-context, and intra-memory conflict. The harm is benchmarked: ClashEval (Wu, Wu & Zou, NeurIPS 2024 Datasets & Benchmarks), 1,200+ questions with calibrated perturbations, finds models override correct priors >60% of the time; WikiContradict (Hou et al., IBM Research, NeurIPS 2024 D&B), 253 human-annotated contradictory-passage instances, finds models fail to represent the conflict; ConflictBank (Su et al., NeurIPS 2024) reports degradation exceeding 20% under noise and far higher under targeted misinformation; MAGIC (Lee et al., EMNLP 2025 Findings) and RAMDocs / MADAM-RAG (Wang et al., COLM 2025) push toward inter-context conflict jointly with ambiguity and noise. Our conflict scenarios are a small, controlled instance of this well-studied failure mode, seeded adversarially so recency and correctness disagree.

Position and recency bias. Liu et al., *Lost in the Middle* (TACL 2024), establish the U-shaped accuracy curve, models use information best at the beginning and end of context. In a memory store surfaced newest-first, this positional effect is a recency effect, which is exactly the tiebreaker a provenance-blind store falls back on, and exactly the one our adversarial seeding turns against it.

Trust-weighted resolution (the closest prior art). RA-RAG (Hwang et al., *Retrieval-Augmented Generation with Estimation of Source Reliability*, EMNLP 2025) resolves conflicting evidence by weighted majority vote, with weights an *unsupervised estimate* of each source's reliability, no ground-truth labels. Credibility-aware generation (Pan et al., 2024) similarly conditions on source credibility. Trust-Align (Song et al., ICLR 2025 Oral) targets *faithful attribution* and evidence- conditioned refusal. These establish that resolving by source trust rather than by surface features is a known and effective idea. Our departure is threefold: the trust signal is a discrete tier fixed at write time, not a continuous estimate computed at read time; it lives in an agent-memory store, not a retrieval pipeline over external documents; and it must be unforgeable by the agent, because in memory the agent is itself a writer.

Agent memory. MemGPT / Letta (Packer et al., 2023) manage memory as OS-style hierarchical tiers (main, recall, archival) to extend effective context; the tiers encode recency and relevance for *paging*, not source trust for *resolution*. Vector-store and summarization memories share this shape. This confirms the gap we target: these systems solve recall, and leave trust-based resolution unaddressed.

Faithfulness vs. truthfulness. A recurring distinction, sharpened in the attribution and summarization-faithfulness literature, separates *faithful* (the output is grounded in its source) from *true* (the source is correct). Our *source-verified* tier is a faithfulness guarantee, not a truth guarantee: it certifies that a value was extracted verbatim from a real message, not that the message was honest. §3.4 makes this boundary explicit, and it is why *human-decided* sits above *source-verified*.

"Context integrity." The phrase appears in adjacent security and privacy literature (e.g. contextual-integrity models of information flow), but not, in the work we surveyed, as a named property of an agent's working context defined by verification, attribution, and resolvability. We use it in that sense and treat this paper as a measurement of one of its components.

---

3. The provenance-tier model

3.1 Facts, tiers, and stores

A fact is a value together with metadata: the subject it concerns, when it was written, and its provenance tier `t in {agent, email, human}` with the strict order

```
human > email > agent
```

read as: a decision made through the owner's authenticated channel outranks a value verified against a source document, which outranks an agent's unverified assertion.

A store answers a question about a subject by selecting one value from the facts about that subject. We compare two selection rules:

- Flat store (`R_recency`): no tier is kept; the only discriminator is time, so it returns the most recent fact.
- Provenance store (`R_tier`): returns the fact of highest tier, breaking ties within a tier by recency.

Both rules are total functions of the stored facts, twenty lines of code, *no model*. This matters: it makes the resolution behavior a property of the *store*, not of any prompt given to an agent, which is what lets §4's result be exact rather than empirical.

3.2 The adversarial seeding assumption

Our conflicts are seeded so that the correct value is not the most recent. Concretely, in each conflict the highest-tier fact (correct) is the *oldest*, and each lower tier is *more recent* and *wrong*. This is the worst case for a recency store, and we argue it is also the realistic one: an agent's fresh guess is written *after* the older verified email it contradicts; a human decision, made once, ages while the agent keeps generating. The whole point of provenance is to override recency with authority, so a fair test must be one where recency and authority disagree. We report this assumption openly; under a friendlier distribution where the newest fact is often correct, the recency store does better, and the gap narrows accordingly.

3.3 Why the tier must be channel-derived (unforgeability)

The order is only meaningful if the top tier cannot be claimed by a writer who has not earned it. In an agent-memory setting the agent is a writer, so a naive design, "let the caller state the tier", is defeated the moment an adversarial or merely mistaken agent labels its own guess *human*.

We therefore derive the tier from the channel, not the claim. Writes arriving through the owner's authenticated surface (an API key the owner holds, or a signed-in dashboard) may mint *human*; writes arriving through the agent surface are *agent*, always, regardless of what the caller asserts. An agent relaying "my operator told me X" may pass that claim along, it is preserved as an *agent-reported attribution*, but it does not change the tier. In the studied system this is enforced at the API boundary; we verified adversarially that an agent call requesting the *human* tier is recorded as *agent* while the human claim is retained as data, and that the value never enters the *human* partition of the store.

This is the paper's systems contribution, and it is the property that retrieval-time trust estimation does not need and does not have: RA-RAG scores *external* sources, none of which is the agent, so it never faces a writer with an incentive to forge its own trust.

3.4 What each tier certifies (faithfulness, not truth)

- *source-verified* certifies faithfulness: the value appears verbatim in a real source message (in the studied system, enforced by a verbatim extraction guard that drops any value not present in the source). It does *not* certify the message was honest; a source that states a wrong price yields a faithfully-recorded wrong value.
- *human-decided* certifies an authenticated human choice, which is why it outranks verification: a human resolving "use the billing email figure" is exercising judgment a source document cannot.

- *agent-asserted* certifies nothing beyond that an agent said it.

The tiers thus order *kinds of evidence*, not *degrees of probable truth*; they are a governance structure, and their correctness under conflict (§4) follows from the seeding, in which the higher tier is the ground truth by construction, mirroring the real relationship where a human decision and a verified source outrank an agent's guess.

3.5 Implementation independence

Nothing above is specific to a product. The design is a set of requirements any memory system can meet:

1. A write-time provenance field on every stored fact, drawn from a small ordered set of tiers. (Categorical; one column, or one key in a serialized record.)
2. Channel-scoped minting: the surface a write arrives on determines the maximum tier it may claim. The privileged tier is reachable only from an authenticated principal's channel; the agent's channel is capped below it. A claim to a higher tier is retained as data but does not raise the field.
3. A tier-first resolution rule at read time: among facts about one subject, return the highest tier, breaking ties by recency.
4. A faithfulness guarantee for the middle tier (optional but recommended): the *source-verified* tier should be minted only when the value is checkable against a stored source, e.g. by a verbatim extraction check.

A researcher could implement all four over any store, a relational table, a document DB, even a JSON file, in well under a hundred lines, and reproduce §4 exactly and §5 with any inference endpoint. The reference implementation we verify against (the Inbin event store, whose write surfaces are an authenticated owner API/dashboard versus an MCP agent plane, and whose *source-verified* tier is enforced by a verbatim guard) is one instantiation of these requirements, not a dependency of the result. The accompanying harness (§Reproducibility) constructs the conflict sets and both resolvers directly, with no product API in the loop.

---

4. Two accuracy models, derived then measured

We model the accuracy of each store as a function of two independent properties of a conflict set, and confirm each derivation with an exact simulation (no model in the loop).

4.1 Conflict depth

A depth-**d** conflict presents `d` competing values across `d` distinct tiers, exactly one correct (the highest tier). Under the adversarial seeding (§3.2), the newest value is always a wrong lower-tier value.

- `R_recency` selects the newest, which is wrong: **`A_flat(d) = 0`** for all `d`.
- A hypothetical fair random tiebreak would select uniformly among `d` values: **`A_random(d) = 1/d`**, given as a friendlier baseline.

- `R_tier` selects the highest tier, which is correct: **`A_tier(d) = 1`** for all `d`.

Measured (K = 50 conflicts per depth, d in {2,3,4}): flat 0% at every depth, random tracking 1/d (50%, 33%, 25%), provenance 100% at every depth, matching the derivation exactly.

4.2 Conflict density

Take a store of `N` facts of which a fraction `ρ` are in (depth-2) conflict; the remaining `1−ρ` are uncontested and both stores answer them correctly. Over `N` questions:

```
A_flat(ρ) = (1 − ρ)·1 + ρ·0 = 1 − ρ
A_tier(ρ) = (1 − ρ)·1 + ρ·1 = 1
```

A flat store's accuracy degrades linearly with how much of its memory conflicts; a provenance store stays at 1. *Measured* (N = 40, ρ in {0, 0.1, 0.25, 0.5, 0.75, 1.0}): every point lands on its derived line, flat at 100, 90, 75, 50, 25, 0%, provenance flat at 100%.

4.3 Cost

The tier is a single categorical field. In a text-serialized store the tag adds on the order of ten tokens per conflicting fact, negligible against the correctness it recovers; in a typed store it is one column and one index. The resolution rule is `O(1)` in the number of tiers.

These closed forms are a scaffold, not a claim to be surprised by. By construction, a recency-only store cannot resolve a conflict whose correct answer is not the most recent, so `A_flat = 0` under adversarial depth and `A_flat = 1 − ρ` under density are near-definitional once the seeding is fixed; a reviewer could fairly say the winning condition was defined to favor the tier resolver. That objection is correct, and it is the point: the deterministic model exists to fix the axis and the extremes cleanly, not to demonstrate anything a careful reader would not already grant. The substantive question, which no construction can settle, is whether *real instruction-tuned models*, given a fair prompt and no tier, are actually pulled toward the recent wrong value, or whether they reason around it. That is an empirical question, and §5 is where the paper's evidence lives.

---

5. Empirical panel

5.1 Setup

Conflicts are programmatically generated, not hand-authored: a fixed seed draws from subject templates (invoice totals, delivery dates, plan tiers, seat counts, contract dates, renewal discounts) crossed with value generators and sentence phrasings, producing a set of 300 conflicts at depths 2, 3, and 4. Values are semantically neutral, nothing in a value string signals which is correct, so a model cannot guess from surface form; the only evidence of correctness is the tier, present in the provenance condition and absent in the flat one. Each conflict is seeded adversarially (§3.2), and 20% are tier-inverted (§5.3). Full generation is reproducible from the seed.

Each conflict is posed to a model twice:

- Flat condition: facts listed with timestamps but no tiers, under a *fair* system prompt instructing the model to give its single best answer and prefer facts that look official or authoritative (its best honest shot without provenance).
- Provenance condition: identical facts, each prefixed with its tier and a system prompt defining the order.

To characterize prompt sensitivity we run three phrasings of each condition's system prompt and pool them, and we report mean accuracy on the conflicting questions with a 95% confidence interval (normal approximation) over trials × phrasings × conflicts, at temperature 0.4, for five instruction-tuned models spanning three architecture families on a common inference provider.

5.2 Results

Model	Flat (mean ± 95% CI)	Provenance (mean ± 95% CI)	Gap
llama-3.3-70b	2% ± 2	45% ± 10	+43
llama-3.1-8b	3% ± 3	36% ± 7	+33
gpt-oss-120b	0% ± 0	0% ± 0	0
gpt-oss-20b	0% ± 0	1% ± 2	+1
qwen3.6-27b*	10% ± 19	10% ± 19	0

* qwen completed only a partial run (rate limiting); its wide interval reflects undersampling and we do not weight it.

The panel separates cleanly into models that engage the resolution task and models that do not. The two engaged models (Llama-70b, Llama-8b) show a large gap that is stable across prompt phrasings: +43 and +33 points, with the provenance interval well clear of the flat interval, and the flat condition near zero, consistent with adversarial seeding turning recency against correctness. The remaining three under-engage: the two GPT-OSS models score near zero in *both* conditions (their outputs refuse or over-format rather than commit to a value, a task-engagement floor, not a provenance effect), and the undersampled qwen shows no separable difference. Across all five, the provenance condition never underperformed the flat condition; we state this as an observation over the tested models, not a universal law.

We note an honest correction against a smaller pilot: an earlier six- conflict run reported a very large qwen gap (+73); at 300 generated conflicts across pooled prompts that separation does not reproduce. Small hand-seeded sets overstate per-model magnitudes; the larger generated set is the number to trust, and it says the effect is real and large *on the models that resolve*, and absent (not reversed) on those that do not.

The reading we support is therefore modest and precise: the direction is a property of the data (the flat condition lacks the information to resolve trust, so it cannot systematically beat the provenance

condition), while the magnitude is a property of the model (whether, and how reliably, it reads and applies the tier at all).

5.3 Where the tier resolver fails (tier inversion)

The deterministic resolver is not magic, and a fair study must measure its failure mode, not only its success. In tier-inverted conflicts (20% of the generated set) the correct value sits at a *lower* tier than a wrong higher-tier fact, modeling a stale or mistaken human decision that outranks a correct verified email. On these:

Subset	Flat	Tier resolver
Non-inverted (n=245)	0%	100%
Inverted (n=55)	0%	0%
All (n=300)	0%	81.7%

The tier resolver inherits the correctness of its ordering: when the tier order reflects ground truth it is exactly right, and when a higher tier is wrong it is *confidently* wrong, exactly as trusting authority should behave. This bounds the claim precisely, provenance tiers convert "which fact is newest" into "which authority is highest," which is an improvement only to the extent the authority ranking is sound, and the improvement is 81.7% at a 20% inversion rate, degrading linearly with inversion. It is a governance mechanism, not an oracle (§3.4, §7).

5.4 Illustrative transcript

One conflict verbatim (Llama-70b; subject "seat count"; correct value from the *human* tier is `48`, the most recent wrong value from the *agent* tier is `36`):

```
flat store       → 36
provenance store → 48
```

Same question, same facts. The flat condition returns the recent wrong value; the provenance condition reads the tier and returns the correct one. This is why we place the deterministic model, not the empirical gap, as the headline: the structural claim, *the flat store lacks the information to resolve trust*, is exact, and the panel shows real agents are pulled by that missing information.

---

6. What is and is not novel

We state this plainly because a reader from the RAG community will, and should, ask.

Not novel. That conflicting evidence harms LLM correctness (ClashEval, WikiContradict, ConflictBank). That models resolve by surface features rather than trust (Xie et al.). That position/recency

biases long-context use (Liu et al.). That resolving by *source trust* beats naive resolution, this is precisely RA-RAG and credibility-aware generation. A reviewer will correctly note that "resolve conflicts by trust rather than recency" is, in spirit, prior art.

Our contribution, stated conservatively. We do not claim to introduce trust-based resolution. We claim that *in current agent-memory designs*, the following combination is not yet standard and measurably improves conflict-resolution accuracy at negligible storage overhead: (i) a discrete tier fixed at write time, rather than a continuous reliability *estimated at read time*; (ii) placed inside an agent-memory store, a setting the trust-RAG methods do not target and the memory systems (MemGPT/Letta) do not address for resolution; and (iii) with the top tier unforgeable by the writing agent, a property that only arises once the agent is itself a source, and that external-source trust estimation therefore never needs. The value is a bridge, carrying a known retrieval idea into agent memory and adding the write-time, unforgeable structure the memory setting demands, plus a measurement of what it buys and where it breaks.

The program this belongs to. We treat conflict resolution as one component of agent context integrity, and read the components as stages an agent's information passes through:

```
communicate  →  store  →  retrieve/resolve  →  act
(reference vs   (write-time    (this paper:        (verified context
 payload)        provenance)    tier resolution)    → tool use)
```

A companion measurement studies the *communicate* stage (reference-based context acquisition); *this* paper studies *resolve*; verification and execution integrity are natural further components. Framing them under one property is what lets each measurement reinforce the others rather than stand as an isolated engineering note. We are careful to present this as our organizing terminology, not a claim to have founded a discipline.

---

7. Limitations

- Seeded and synthetic. Six subject templates crossed with value generators and phrasings, expanded to 300 programmatically generated conflicts under adversarial seeding. This is a controlled demonstration, not a field study; the density and depth results are exact *given the construction*, and are close to definitional once the seeding is fixed. The value is that the construction is transparent and reproducible, not that it samples a natural distribution.
- The deterministic result is near-tautological by design. A recency-only store cannot, in principle, resolve a conflict whose correct answer is not the most recent. We present it not as a surprise but as the exact statement of a structural fact, and lean on §5 for the non-obvious part.
- Empirical panel is narrow in models, if not in items. The deterministic and conflict sets are large (300 generated conflicts) and the prompt sweep is real (three phrasings), but only two of five models engaged the resolution task at all; the effect, though large and prompt-stable on those two, rests on $n = 2$ *engaged* models on a single inference provider. More providers and models

that reliably follow instructions are the clear next step; the current evidence supports "a large effect where models resolve, no reversal where they do not," not a population-level magnitude.

- The gain is bounded by the tier ordering's own correctness (§5.3). We measure this directly via tier inversion: on the 20% of conflicts where a higher tier is wrong, the resolver is 0% correct, dropping the overall figure to 81.7%. The tier resolver is a governance mechanism, not an oracle, and its benefit degrades linearly with the rate at which the authority ranking is itself wrong. A deployment whose *human* tier is frequently stale would see a correspondingly smaller gain; keeping that tier trustworthy is a precondition, not an assumption we hide.
- Prompt sensitivity is characterized but not exhausted. We pool three phrasings per condition (§5.1) and the engaged-model gap is stable across them, but three phrasings is a spot check, not a systematic robustness study. The structural argument (§3.4) is that no flat prompt can recover information the store never kept; the empirical *magnitude* remains prompt-dependent to a degree we bound only coarsely.
- Governance, not truth. The tiers order kinds of evidence. A *human-decided* fact can still be a human's mistake; the model guarantees the store resolves *to the tier we defined as authoritative*, which is a governance guarantee, not an oracle.
- Adversary model is narrow. Unforgeability covers an agent forging its own tier via the API; it does not cover a compromised owner credential, which by construction speaks with owner authority.

---

8. Conclusion

The reliability of an agent's answers is usually framed as a recall problem, remember the right fact. We have argued that a second problem, *resolution*, when remembered facts conflict, is decided not by the model but by what the store chose to keep, and that keeping a fact's provenance as a discrete, write-time, unforgeable tier lets the store resolve conflicts by authority rather than by recency. Two closed-form models bound the extremes under an adversarial construction, and, more importantly, a five-model panel shows real instruction-tuned agents are in fact pulled toward the recent wrong value when the tier is withheld, with a direction that holds across the panel and a magnitude that depends on the model. The idea of trust-based resolution is not ours; what we offer is the evidence that, in current agent-memory designs, moving it into memory as an unforgeable write-time tier measurably improves conflict-resolution accuracy at negligible storage overhead, and a characterization of where that improvement fails (tier inversion). We read this as one component of agent context integrity, the property that information an agent communicates, stores, retrieves, and acts upon stays reliable across all four stages, and offer it as one falsifiable measurement in a program of them.

Reproducibility

The conflict experiment is a single self-contained script. The deterministic models run with no external dependency and produce the exact depth and density tables on every run; the empirical panel addition-

ally requires an inference API key and reports the panel table and a verbatim transcript.

```
DETERMINISTIC_ONLY=1 node scripts/research-conflict-study.mjs   # exact models
GROQ_API_KEY=... node scripts/research-conflict-study.mjs         # + panel
```

Deterministic output is identical across runs; the empirical panel varies by model and provider availability and is reported with per-model mean ± 95% confidence interval.

References

Selected; full survey retained with the source material.

- Xu et al. *Knowledge Conflicts for LLMs: A Survey.* EMNLP 2024.
- Wu, Wu & Zou. *ClashEval: Quantifying the tug-of-war between an LLM's internal prior and external evidence.* NeurIPS 2024 D&B.
- Hou et al. *WikiContradict: A Benchmark for Evaluating LLMs on Real-World Knowledge Conflicts from Wikipedia.* NeurIPS 2024 D&B.
- Su et al. *ConflictBank: A Benchmark for Evaluating Knowledge Conflicts in LLMs.* NeurIPS 2024.
- Liu et al. *Lost in the Middle: How Language Models Use Long Contexts.* TACL 2024.
- Xie et al. *Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of LLMs in Knowledge Conflicts.* ICLR 2024.
- Hwang et al. *Retrieval-Augmented Generation with Estimation of Source Reliability (RA-RAG).* EMNLP 2025.
- Song et al. *Measuring and Enhancing Trustworthiness of LLMs in RAG through Grounded Attributions and Learning to Refuse (Trust-Align).* ICLR 2025 (Oral).
- Wang et al. *Retrieval-Augmented Generation with Conflicting Evidence (RAMDocs / MADAM-RAG).* COLM 2025.
- Lee et al. *MAGIC: Inter-Context Conflict benchmark.* EMNLP 2025 Findings.
- Packer et al. *MemGPT: Towards LLMs as Operating Systems.* 2023 (Letta).