yAI SYSTEMS NOTE 003

When Retrieved Evidence Conflicts with Model Memory: Knowledge Conflicts in RAG Systems

Version 1.0 March 26, 2026

Yassin Hafid
Founder & CEO

LinkedIn

ABSTRACT

Retrieval-augmented generation assumes that models will prioritize retrieved evidence over their internal parametric memory. Empirical studies show this assumption frequently fails; indeed, models exhibit confirmation bias, selectively absorb evidence that aligns with prior beliefs, and paradoxically resist updating knowledge that changes most often across studied model families. This failure is most acute precisely where retrieval matters most: in high-dynamicity domains such as financial markets, regulatory environments, and time-sensitive data, where retrieved evidence is most likely to conflict with outdated parametric priors. This note examines why knowledge conflicts between retrieved context and parametric memory undermine the reliability of RAG systems, even when retrieval is successful.

1. The Question

If a RAG system retrieves a passage that contradicts what the model already "knows," which source wins? And does the model signal when a conflict has occurred?

---

2. Scope and Definitions

> **Parametric Knowledge**
>
> Facts encoded in model weights during pretraining.
>
> **Contextual Knowledge**
>
> Information supplied at inference time via retrieved passages.
>
> **Knowledge Conflict**
>
> A discrepancy between parametric knowledge and retrieved contextual knowledge.
>
> **Memorization Ratio**
>
> The proportion of responses in which a model defaults to its parametric answer despite conflicting retrieved evidence.[2]
>
> **Confirmation Bias**
>
> The tendency of a model to accept external evidence more readily when it partially aligns with parametric memory, even when the overall evidence conflicts.[1]

This note focuses on context-memory conflicts in RAG settings where retrieved passages are factually inconsistent with model priors.

---

3. Key Findings

- **LLMs exhibit contradictory behavior under conflict:** They can be receptive to conflicting evidence when it is coherent and convincing; however, they show strong confirmation bias when external evidence partially aligns with parametric memory while other parts conflict.[1]

- **Context-receptivity dominates single-evidence scenarios:** Memorization ratios generally remain below 50% across 12 LLM instances spanning four model families; this indicates that models are broadly receptive when presented with a single piece of conflicting evidence. However, this resistance is not uniform: models exhibit their highest memorization ratios in misinformation conflicts, showing a stronger tendency to default to parametric memory when faced with counter-factual claims. Conversely, they demonstrate lower memorization in temporal and semantic conflicts, indicating they are relatively more willing to adopt context that reflects a change in time or a shift in meaning than they are to adopt outright misinformation.[2]

- **Synthetic benchmarks overstate context utilisation:** Comparisons between real-world retrieval and synthetic datasets show that the latter provide inflated context utilization results.[3] In realistic settings, model receptivity is substantially lower due to the complexity and noise of retrieved evidence. Interestingly, while some models exhibit negative ACU scores (actively moving away from the evidence), this "context-repulsion" is actually exaggerated in synthetic data and is rarely observed in realistic retrieval scenarios.[3]

- **Fact dynamicity predicts conflict failure, not fact popularity:** Fact dynamicity (i.e., the frequency with which a fact changes) is the strongest negative predictor of successful context utilization, outperforming fact popularity. Models are more resistant to updating knowledge about dynamic facts than static facts, demonstrating a "stability bias" that prioritizes parametric memory for information that changes most often in the real world.[4]

- **Fragility Under Knowledge Shift:** RAG systems suffer a substantial performance drop when retrieved documents reflect updated or hypothetical knowledge that conflicts with parametric priors. The failure is most acute in tasks requiring the seamless integration of contextual and parametric knowledge, specifically "Distant Implicit" questions; this highlights a fundamental fragility in how models combine their internal memory with shifting external evidence.[5]

---

4. Technical Deep Dive: Why Conflicts Go Unresolved

A. Confirmation Bias and Coherence Sensitivity

Model behavior under conflict is not uniform. Evidence presented coherently and convincingly, rather than through simple entity substitution, is far more likely to override parametric memory.[1] This has a practical implication: low-quality or poorly phrased retrieved passages may fail to displace incorrect parametric priors even when factually correct.

B. The Paradox of Dynamic Facts

Counter-intuitively, models are most resistant to updating the very facts that change most often. While one would expect a model to defer to context for shifting information, fact dynamicity is actually the strongest predictor of context-utilization failure outperforming even fact popularity. This is driven by intra-memory conflict; one possible explanation is that dynamic facts appear in multiple conflicting states throughout pre-training corpora, producing entangled internal representations that are harder to displace than the cleaner representations associated with static facts.

C. Benchmark Inflation and Real-World Performance

Most prior studies of knowledge conflict used synthetic datasets that do not reflect the complexity of real retrieved contexts. When evaluated on real-world retrieval contexts, models show substantially lower context utilisation, and no single context characteristic (e.g., source authority, semantic similarity) reliably predicts when a model will or will not follow the evidence.[3] This means conflict-resolution performance reported in controlled settings may not transfer to deployed systems.

---

5. Practical Taxonomy of Conflict Failure Modes

- **Coherence-Gated Override:** Model ignores retrieved evidence when it is not presented coherently enough, even if factually correct.[1]
- **Confirmation Absorption:** Model accepts evidence that partially supports prior beliefs while discarding the conflicting portion.[1]
- **Dynamic Fact Resistance:** Model fails to update knowledge about temporally or contextually variable facts, precisely the facts where retrieval is most needed.[4]
- **Knowledge Shift Collapse:** RAG performance degrades substantially when retrieved documents reflect updated knowledge that contradicts training-time priors.[5]
- **Benchmark-Reality Gap:** Context utilisation measured on synthetic conflict data does not generalize to real-world retrieval; models appear more robust in controlled settings than in deployment.[3]

---

6. Implications for AI System Design

- **Do not assume conflict detection is implicit:** Models do not reliably signal when retrieved evidence conflicts with parametric memory; explicit conflict detection mechanisms are needed.
- **Coherence of retrieved passages matters beyond relevance:** A retrieved passage may be topically relevant but insufficiently coherent to override prior beliefs; retrieval systems should optimize for coherence as well as relevance.[1]
- **Evaluate on real-world retrieval contexts:** Synthetic conflict benchmarks overstate context utilisation; system evaluation should use real retrieved data with realistic stance diversity.[3]
- **Prioritize temporal knowledge as a high-risk conflict category:** Dynamic and temporally variable facts represent the highest-risk knowledge conflict scenario and warrant targeted retrieval and verification strategies.[4]

---

7. Open Questions

- **Conflict detection:** Can models be trained or prompted to reliably detect when retrieved evidence conflicts with parametric memory, and to signal uncertainty accordingly?
- **Resolution criteria:** When should a model defer to retrieved evidence and when to parametric memory? No principled resolution criterion currently exists that generalizes across conflict types.[2]
- **Coherence and trust:** What properties of retrieved passages, beyond topical relevance, determine whether a model will treat them as authoritative?[1][3]
- **Dynamic knowledge pipelines:** How should RAG systems be designed when the target domain involves high fact dynamicity, such as financial markets or regulatory environments?

The above findings also raise forward-looking engineering questions not yet addressed in the literature, to the best of our knowledge:

- **The Internal Suppression Hypothesis:** Can we use "Activation Steering" to proactively silence conflicting parametric weights when a high-dynamicity fact is detected?
- **Meta-Cognitive Conflict Detection:** Can a model be trained to output a "Conflict Internalization" signal before decoding, flagging the tension between memory and context?
- **Multi-Hop Conflict Propagation:** If a model accepts a "Distant" fact in step 1 of a reasoning chain, does its "Acceptance" degrade or strengthen as it moves through subsequent implicit reasoning steps?

---

8. References

[1] Xie et al., Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts, ICLR 2024 (arXiv:2305.13300, 2024).

[2] Su et al., ConflictBank: A Benchmark for Evaluating the Influence of Knowledge Conflicts in LLMs, NeurIPS 2024.

[3] Hagström et al., A Reality Check on Context Utilisation for Retrieval-Augmented Generation, ACL 2025.

[4] Augenstein et al., Understanding the Interplay between LLMs' Utilisation of Parametric and Contextual Knowledge: a keynote ECIR 2025, arXiv:2603.09654, 2026.

[5] Zheng et al., KnowShiftQA: How Robust are RAG Systems when Textbook Knowledge Shifts in K-12 Education?, ACL 2025.

This note synthesizes findings from recent research on knowledge conflicts in RAG systems. The interpretations presented reflect the author's reading of the current literature.