RAG Persistence Vulnerability: Full Technical Disclosure

Author: Reamond Lopez (@Aequitas_Architech)
Date: February 2026
Status: Disclosed under Google VRP, triaged and mitigated by vendor
Framework: VERITAS Evaluation Suite v3.1.0-GOLD

Executive Summary

This disclosure documents a verified persistence vulnerability class in Retrieval-Augmented Generation (RAG) systems.

Across 43 certified evaluation runs using Google's Gemini API, 0% of tested reset mechanisms successfully returned the model to a verified clean baseline.

Google accepted the finding under its Vulnerability Reward Program, implemented mitigations, and closed the case. This document is published with vendor approval.

Core Finding

Retrieved content influence persists beyond intended session boundaries, surviving documented reset mechanisms.

The issue is architectural rather than prompt-based or adversarial in nature.

Implications include privacy boundary violations, stale knowledge persistence, and compliance risks.

VERITAS Evaluation Methodology

VERITAS establishes a frozen baseline, introduces controlled influence, applies reset mechanisms, and measures deviation.

All runs were deterministic, reproducible, and cryptographically sealed.

The Trinity framework (Mind, Sword, Shield) governed logical soundness, measurement precision, and containment.

Certified Results Summary

Total certified runs: 43

Clean resets achieved: 0%

All documented reset mechanisms failed under controlled conditions.

Responsible Disclosure Timeline

Discovery: January 2026

VRP Submission: January 15, 2026

Mitigation Implemented: February 5, 2026

Case Closure: February 10, 2026

Public Disclosure: February 16, 2026

Conclusion

RAG reset isolation assumptions are fragile under normal operating conditions.

This vulnerability class extends beyond a single vendor implementation.

The findings motivate architectural changes and standardized reset integrity evaluation.