Audit-Grade Evaluation of RAG Reset Robustness

Purpose
This document describes a deterministic measurement framework used to evaluate whether common reset procedures in retrieval-augmented language model systems return behavior to a verified clean baseline after ingestion of untrusted retrieved content.

Key Observation
Across controlled runs, lexical residue appears only after ingestion of untrusted retrieved content and persists across reset mechanisms including thread isolation, context flushing, and cooldown-based reinitialization. Clean baseline runs remain clean.

Method Overview
- Define clean vs contaminated baselines
- Apply standard reset / isolation procedures
- Compare outputs statistically, not semantically
- Detect short lexical signatures that persist across runs

Design Constraints
This appendix omits prompts, payloads, reproduction scripts, and vendor-specific details. The focus is on measurement validity and infrastructure requirements, not exploitation.

Infrastructure Note
Consumer-grade hardware introduces scheduler and I/O nondeterminism during long-horizon testing. Dedicated local compute is required to preserve deterministic artifacts during extended evaluation runs.