Deterministic Sovereign RAG

Source: Publicaciones · **Status:** Shipped · **Period:** 2025-08 (active)

Academic publication covering byte-exact, local-first verification and retrieval primitives for auditable RAG workflows.

The challenge

Most retrieval-augmented generation pipelines are non-deterministic across re-runs, so an audit cannot trace which evidence supported which generated claim. The 'retrieval' half of RAG is a black box.

My role

I published the math foundation for O(1) local context compaction: FNV-1a 128-dim projection, L2 normalization, cosine similarity, with Chernoff-Hoeffding concentration bounds. The work anchors a companion repository and a meta-tool.

What I did (3–5 decisions)

1. Deterministic projection of context windows

Why: A 128-dimensional FNV-1a projection collapses a context window to a fixed-size vector that is reproducible across runs. The projection is O(1) per token and round-trip safe.

Tradeoff: FNV-1a is not cryptographically secure; the projection is for similarity scoring, not for tamper resistance. Cryptographic anchoring is a separate layer.

2. Concentration bounds make the similarity claim auditable

Why: A cosine similarity above a threshold is not by itself a claim; the Chernoff-Hoeffding bound makes it a claim with a confidence interval. Auditors get a number, not a vibe.

Tradeoff: The bounds are conservative — a relevant document that would pass a looser threshold can fail the bound. The discipline is the point.

3. Local-first execution end to end

Why: The projection runs in-process; the embeddings never leave the operator's machine. The sovereign-rag posture is the design constraint, not a marketing line.

Tradeoff: CPU cost is on the operator. The cost is the price of an audit chain that does not depend on a third-party embedding service.

What changed

Context compaction

Baseline: Variable window, non-reproducible

Fixed 128-dim projection, deterministic

Audit confidence on a similarity claim

Baseline: Vibes, no bound

Chernoff-Hoeffding confidence interval

Operator-controlled execution

Baseline: Hosted embedding API

Local-first, in-process

Overall tradeoffs

The deterministic posture costs CPU. A reviewer who does not need the audit chain can use a hosted embedding service and skip the projection entirely.

Lessons

Determinism is not the same as security; determinism is the precondition for an audit chain to exist at all.

Stack

Math | FNV-1a | L2 normalization | Chernoff-Hoeffding

Evidence

- [gist] Deterministic Sovereign RAG — math foundation (EN) — gist.github.com/louzt/d1ce71c05460c2c32bf31342cd0c6a3f