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 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

Stack

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