

Falsifiers, Benchmarks, and Evaluation Plan

Falsifier Suite

F1 Confluence break; F2 Type III mirage; F3 Pal witness failure; F4 Legality hallucination; F5 Energy regression; F6 Retrieval regression.

Bench Plans

A/B cosine RAG vs invariant key CQE RAG on hallucination prone QA; measure legality rate, determinism, energy/commit. Replay audits on residues only. External pin loop demonstrating STOP→NTER→OPEN.