

FLAGSHIP RESEARCH ENGINEERING PROJECT

RAG Observatory

A trace-based toolkit for turning RAG runs into inspectable evidence: stage-aware reports, configuration comparisons, failure labels, and explicit observability gaps.

PUBLIC EVIDENCE

GitHub repository

Live Hugging Face demo

Project overview

What the system records

Query	Retrieval	Reranking	Context	Prompt	Generation	Evaluation	Diagnostics
OBSERVABILITY CONTRACT Maps public traces into ordered stage spans. Accepts offline OTLP/HTTP JSON exports containing OpenInference retriever, reranker, and LLM spans. Large payloads remain outside low-cardinality attributes.				DIAGNOSTIC BOUNDARY Missing stages stay explicit. If an imported trace does not prove which chunks entered the prompt, the adapter records the gap instead of inventing selected context.			

Controlled SciFact retrieval-depth run

Twenty fixed BEIR SciFact test queries were run with deterministic BM25 at retrieval_top_k=1 and 5. The query set, top-1 context policy, extractive generator, and qrels evaluator remained fixed.

Configuration	Retrieval hit	Failure-labelled traces	Retrieval-stage failures	Context failures
top_k=1	13/20 (0.650)	7/20	7/20 (35%)	7/20 (35%)
top_k=5	14/20 (0.700)	20/20	20/20 (100%)	7/20 (35%)

What this exposed

- Retrieval hit improved by one query, but the current noise heuristic labelled every top_k=5 trace because any irrelevant retrieved document triggers the signal.
- The largest observed change localized to retrieval: 35% to 100%. Context and generation failure rates both remained 35%.
- The report therefore reveals both a retrieval-depth trade-off and a limitation in the diagnostic heuristic; it does not convert a smoke experiment into a benchmark claim.

Scope boundary. This is a real retrieval-stage run over 40 public traces, not a full generative RAG benchmark. Twenty sorted queries support an inspectable engineering comparison, not a population-level or causal claim.

Implementation evidence and limits

40	20	107	8
public SciFact traces	fixed test queries	automated tests	explicit RAG stages
IMPLEMENTED Public trace conversion, OTLP/OpenInference JSON import, stage-aware Markdown reports, configuration reports, and per-query failure changes.		NOT CLAIMED A production tracing backend, network receiver, full generative benchmark, dataset-wide result, or causal conclusion beyond the controlled exported runs.	

Reliable mobile system - CiboCompass

A full-stack food-exploration application for international students in Italy, built with React Native/Expo, a Go REST API, and SQLite.

React Native / Expo client	Go REST API	SQLite persistence	Retry and idempotency
----------------------------	-------------	--------------------	-----------------------

- Implemented local caching, a durable pending-feedback queue, and retry on startup, foregrounding, or later submissions.
- Reused idempotency keys across retries and collapsed queued updates to the newest rating per dish and nationality.
- Documented the system boundary explicitly: offline delivery is handled, but the project does not claim complete multi-device synchronization.

OFFLINE BEHAVIOR Ratings remain in a durable local queue when the network is unavailable. Retries run on application startup, foregrounding, and later submissions.	WRITE SAFETY A retry keeps the same idempotency key, while queued updates collapse to the newest rating for the same dish and nationality.	EXPLICIT LIMIT The project does not claim actor identity, cross-device conflict resolution, or a complete synchronization protocol.
---	--	---

[GitHub repository](#)

Merged open-source work

MTEB #4861 Documented the model-evaluation request workflow so contributors can submit requests through a clearer, reviewable path. View merged pull request	MTEB Leaderboard #31 Added model language-scope metadata together with tests and accessibility coverage in the leaderboard frontend. View merged pull request
---	--

Client systems delivery - Idearia

DATA TOOL Built a client-facing web tool for inspecting and analysing uploaded Excel data.	OPERATIONS TOOL Built a second client system for managing employee leave requests and administrative records.	WEB AND AUTOMATION Maintained WordPress/PHP sites, created minisites with domain and email setup, delivered two n8n workflows, and managed Xibo media content.
--	---	--

Methods and application context

IR AND NLP BM25, FAISS, cross-encoder reranking, T5, retrieval diagnostics, and RAG evaluation.	EVALUATION Paired bootstrap, confidence intervals, controlled comparisons, failure taxonomies, and qrels-based checks.	SYSTEMS Python, Go, SQL, REST APIs, SQLite, Docker, GitHub Actions, offline queues, and idempotent writes.
Education. B.Sc. Applied Computer Science and Artificial Intelligence, Sapienza University of Rome; expected December 2026.		Target. German and European M.Sc., HiWi, research assistant, and research engineering opportunities in IR, NLP, reliable RAG, and reproducible ML.