Designing Observable Retrieval-Augmented Systems

Research note - portfolio demo document.

A placeholder paper about making RAG systems diagnosable by default.

Demo-safe	Research note	Replaceable asset

ABSTRACT

Abstract

RAG systems fail in ways that are hard to explain if retrieval, ranking, prompting, and evaluation are not legible. Observability should describe not only whether a request passed, but why it behaved the way it did.

OBSERVATION LAYERS

Observation layers

Useful observability crosses product and model boundaries.

- Request traces with stage timing.
- Retrieved document snapshots and ranking features.
- Prompt templates and revision history.
- Outcome scoring tied to user feedback and offline evals.

RECOMMENDED DASHBOARD FIELDS

Recommended dashboard fields

Field	Reason
query + intent	Tells reviewers what the system thought the user wanted.
retrieved context IDs	Makes context inspection possible.
answer confidence or quality flags	Supports triage and escalation.
human override notes	Captures operator learning for later improvement.

PRACTICAL TAKEAWAY

Practical takeaway

The system should produce enough evidence that a reviewer can answer: what did the model see, what did it choose, and what should change next?