RAG Assessment

Benchmark retrieval quality, evidence grounding, citations, and operating efficiency before you redesign the architecture.

ASSESSMENT GOAL

When the answer looks plausible but the failure mode is unclear.

01 Why assess before you rebuild?

Isolate the failure mode before changing the stack.

Assess first when you see

- Weak retrieval — Required evidence is missed or critical
- Evidence misalignment — Claims and citations do not match the
- Operating uncertainty — Latency, cost and repeatability block

PRINCIPLE · Measure the baseline before changing the stack.



02 What do we assess?

Compare systems against the same questions and criteria.

01 Retrieval Quality

Recall@k · Precision@k · Context Relevance

Determine whether the system retrieves the evidence required to answer.

02 Groundedness & Citation

Claim Grounding · Citation Validity

Verify that claims are supported and citations point to the right evidence.

03 Failure Analysis

Retrieval · Evidence · Answer Failure

Separate failure modes so remediation targets the right layer.

04 Efficiency & Consistency

Latency · Token/Cost · Repeatability

Measure repeated - run stability and operating trade - offs.

03 How the assessment works

1 Baseline
Establish the current - state benchmark

2 Compare
Compare configurations like - for - like

3 Verify
Validate claims, citations and evidence

4 Improve
Prioritize remediation and PoC scope

04 Deliverables and validation scope

Assessment Output

- Baseline Scorecard — A measurable current - state benchmark
- Failure Taxonomy — Failure modes across retrieval, evidence and answer
- Prioritized Remediation Plan — Actions ranked by expected impact

RECOMMENDED START Benchmark before you redesign the architecture.

LIMITED VALIDATION SNAPSHOT

RAG Assessment v1.2.1

6/6 Golden6 E2E	22 Fixed Claims	44 Evidence Eval.
38/44 Citation Scope	24/24 Provenance	23/23 Unit Tests

38/44 is not an overall RAG or answer - accuracy score. Human Gold approval and an independent held - out benchmark are still pending; the figures reflect technical validation on a bounded evaluation set.