ILLUSTRATIVE DEMONSTRATION

RAG Evaluation Sprint

A worked example of a retrieval-augmented generation (RAG) evaluation method, from source retrieval to answer quality.

Why evaluation comes before scale

A knowledge assistant can retrieve the wrong source, miss policy exceptions, or cite stale content. This worked example assumes a twenty-question set. The example percentages illustrate a reporting format; no evaluation run is reported.

20	5	2
assumed questions	evaluation dimensions	example configurations

Evaluation dimensions

- Retrieval recall: the required source appears in the candidate set.
- Groundedness: every material claim is supported by retrieved context.
- Citation accuracy: the cited passage actually supports the answer.
- Completeness: the answer includes required exceptions and conditions.
- Abstention: the system declines when the corpus cannot support an answer.

Illustrative evaluation scorecard

Example configuration	Retrieval recall	Grounded answers	Citation accuracy	Completeness	Correct abstention
Baseline: fixed chunks, top-3	70%	65%	75%	Not scored	60%
Revised: structure-aware, reranked top-5	90%	85%	95%	Not scored	90%

Illustrative values only, not measured performance. This bundled change cannot isolate the effects of chunking, reranking, or retrieval depth. Completeness is not scored.

Example failure analysis

Failure	Illustrative pattern	Possible next test
Missed exception	Answer used the general rule but missed a regional exception	Preserve document hierarchy and attach parent headings to chunks
Stale policy	Older duplicate ranked above current policy	Add effective-date metadata and freshness weighting
Unsupported synthesis	Answer combined two passages into an unsupported conclusion	Require claim-level citations and groundedness checks
False confidence	System answered a question absent from the corpus	Add uncertainty threshold and explicit abstention behavior

Define a release gate

- Freeze a versioned evaluation set and run it on every retrieval or prompt change.
- Set minimum thresholds by risk category rather than relying on one average score.
- Log retrieval candidates, citations, latency, token cost, and abstention outcome.
- Route low-confidence and high-impact questions to a person.

Evidence note

Original Axtryll demonstration, wording revised September 19, 2026. The question count and percentages are illustrative assumptions. No published dataset, measured improvement, production readiness, or client result is claimed.