SYNTHETIC DEMONSTRATION

RAG Evaluation Sprint

A compact proof of how Axtryll evaluates retrieval quality, grounded answers, citations, and production-readiness before a knowledge assistant ships.

Why evaluation comes before scale

A knowledge assistant can sound convincing while retrieving the wrong source, missing policy exceptions, or citing stale content. This sample uses a synthetic twenty-question support corpus to demonstrate a repeatable evaluation method.

20	5	2
synthetic test questions	failure categories	configurations compared

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.

Synthetic benchmark results

Configuration	Retrieval recall	Grounded answers	Citation accuracy	Correct abstention
Baseline: fixed chunks, top-3	70%	65%	75%	60%
Revised: structure-aware, reranked top-5	90%	85%	95%	90%

Failure analysis

Failure	Observed pattern	Remediation
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

Production 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

The corpus, questions, and benchmark values are synthetic and were created for this demonstration. They show the evaluation approach and deliverable structure, not a production claim or client result.