Xtensyon.

TECHNICAL BRIEF

Enterprise Evaluations for RAG: Scorecards, Golden Sets, and Release Gates

February 5, 2026 • Xtensyon Labs, Applied AI Team • 7 min read

Most RAG projects fail quietly: the demo works, then accuracy drifts after a few indexing tweaks. This brief introduces a scorecard approach with simple metrics, a golden query set, and release gates to keep enterprise RAG systems stable in production.

TL;DR

- Keep a small golden set you can run often; daily beats perfect.
- Score retrieval and citations, not just generation quality.
- Gate releases so indexing or prompt changes cannot ship with regressions.
- Log retrieval context so failures are actionable, not mysterious.

Executive Summary

Enterprise RAG needs more than “it looks correct.” Teams tweak chunking, swap rerankers, and add data sources, then wonder why users suddenly get confident but wrong answers. The fix is not endless prompt debates. Treat the RAG stack like a production system: define what “good” means, measure it, and block releases that make it worse. This brief outlines an evaluation program we have used to ship RAG updates safely: a golden query set, a one-page scorecard, and release gates that catch regressions before users do.

Why It Matters

Without evaluations, RAG quality becomes opinion-driven. That is risky because the tool can sound authoritative even when the citation is wrong or the content is stale. A basic evaluation loop gives predictability, accountability, and speed. It also becomes shared language between engineering, product, and risk teams, which matters when there are many stakeholders.

What We Built

- A “golden set” of 50–200 representative queries mapped to expected sources, updated monthly as business processes change.
- A single scorecard that tracks citation accuracy, answer completeness, retrieval freshness, and non-answer rate (“I don’t know” when appropriate).
- A regression harness that replays the golden set against each candidate build (indexing changes, ranking changes, prompt changes).
- A lightweight review workflow: failed cases are tagged (stale doc, wrong ACL, chunk mismatch, ranking error, prompt injection) and routed to an owner.

Observed Outcomes

- Faster releases because teams stop guessing which knobs to turn and focus on the failing categories.
- Fewer “silent failures” after content updates by tracking freshness and version-aware citations.
- More reliable stakeholder reporting with a simple dashboard: trends over time, not one-off anecdotal screenshots.

Implementation Notes

- Keep the golden set small enough to run often. Daily is better than perfect. You can expand once the harness is stable.
- Don’t overfit expected answers. Prefer checks like: correct source cited, key entities present, policy constraints respected.
- Log retrieval context in evaluation runs (top-k doc IDs + versions). When a test fails, you’ll know if it’s retrieval or generation.
- Include adversarial queries: prompt injection attempts, ambiguous questions, and permission-sensitive requests.

Governance & Risk

- Define a kill switch: if quality drops past a threshold, roll back indexing/prompt changes automatically.

© 2026 Xtensyon. All rights reserved.

- Maintain evidence. Save evaluation reports per release so you can answer “what changed” during audits or incidents.
- Watch for “helpful but unsafe.” A model can pass factual tests while still leaking restricted content if ACL checks fail.

Release Checklist

- Do we have a golden query set tied to real business workflows?
- Do we measure citation correctness, not just “nice-sounding” answers?
- Can we run regression tests on every indexing/ranking/prompt change?
- Do we triage failures into categories with clear owners?
- Do we have a rollback rule when quality degrades?

Conclusion

If you want RAG to be trusted, you need a boring, repeatable evaluation loop. The good news: you do not need a research team or fancy benchmarks. Start with a golden set, add a scorecard, and enforce release gates. Once it is in place, your team will ship faster because you are no longer relying on gut feel.

Keywords

RAG evaluation · LLM testing · golden dataset · prompt regression · AI quality assurance · hallucination detection · enterprise AI governance