Runbook: Vector Retrieval Tuning (TopK, Filters, Rerank)

Last updated: 2026-01-23

Purpose

Tune retrieval so the LLM answers are grounded and consistent. Applies to FAISS/Chroma/Pinecone/Azure AI Search.

Default Retrieval Settings

Parameter	Default	Notes
top_k	5	Start here; raise to 10 for broad questions.
score_threshold	optional	Use if you see irrelevant chunks.
metadata filters	agent/env/source	Filter to reduce noise.
rerank	optional	Use LLM rerank for better precision.

Symptoms of Poor Retrieval

- LLM cites wrong runbook or mixes environments (dev vs prod).
- Answers are generic and ignore runbook-specific steps.
- Hallucinations when vector hits are weak or unrelated.

Mitigation

- Add metadata filters: agent='lambda' for Lambda questions; env='prod' only in production UI.
- Use a short query rewrite step (LLM) before embedding.
- If PDFs are long, add section-aware chunking using headings.

Rerank Strategy

- Retrieve top_k=20 cheaply, then rerank to top_k=5 using a lightweight reranker.
- Always pass the reranked snippets into the final answer with citations (source + section).