Contents

CRAG-Based LMKR Assistant

1. Introduction

This document presents a fully disciplined implementation of Corrective Retrieval-Augmented Generation (CRAG) for an LMKR-focused conversational assistant. The system is designed to ensure factual accuracy, minimize hallucinations, and correctly handle both static and time-sensitive queries by applying CRAG principles rigorously.

2. Motivation

Traditional RAG systems often overuse web search, mix internal and external evidence, and generate answers from incomplete context. CRAG addresses these shortcomings by introducing a grading mechanism that evaluates the sufficiency of retrieved knowledge before generation.

3. System Architecture Overview

The system follows a strict CRAG flow:

1) Internal retrieval using FAISS
2) Document grading (gatekeeper)
3) Conditional corrective retrieval (query rewrite + web search)
4) Evidence-grounded generation

4. Internal Retrieval (FAISS)

Static LMKR documents are embedded using a transformer-based embedding model and indexed using FAISS. FAISS acts as the primary and trusted knowledge source. Retrieval is evidence-only and does not influence routing decisions.

5. Document Grading – CRAG Gatekeeper

The grading component evaluates whether retrieved documents can answer the user query completely and confidently. It is the sole authority that decides whether corrective retrieval is required. If grading fails, internal documents are discarded.

6. Corrective Query Rewriting

When grading determines that internal documents are insufficient, the system rewrites the query to improve external retrieval. Query rewriting is only performed in this corrective scenario and never influences internal retrieval.

7. External Retrieval (Web Search & BambooHR)

External retrieval is used strictly as a corrective mechanism. Web search is restricted to official LMKR domains. Career-related queries extract live job data from LMKR's BambooHR portal to avoid hallucinations.

8. Evidence-Grounded Generation

The generation component produces answers using only the currently accepted evidence. If evidence is insufficient, the system responds that it cannot confirm the answer rather than guessing.

9. CRAG Discipline Enforcement

The system enforces CRAG discipline through strict invariants:

- Retrieval never decides routing
- Grading is the only gatekeeper
- Web search is corrective, not augmentative
- Generation occurs exactly once
- Evidence sources are never mixed

10. Conclusion

This project demonstrates a disciplined and production-ready CRAG system. By enforcing strict evidence control, it delivers accurate, consistent, and trustworthy responses for LMKR-related queries. The latency for this falls to under 10s on average cases.