# RAG System Design & Evaluation (Consolidated)

## Overview

This document consolidates concepts from the base and enriched RAG docs. It covers retriever design, prompt augmentation, grounding evaluation, latency tuning, and governance best-practices.

## Retriever Fusion Strategies

Combine lexical and semantic search using Reciprocal Rank Fusion (RRF) and weighted score normalization. When corpora include code or highly structured text, consider BM25→Dense re-ranking with Cross-Encoder rerankers.

## Knowledge Graph Integration

Augment retrieval by linking entities and relations from a Knowledge Graph (KG). Use KG hops to expand recall and fetch related context blocks. Store KG embeddings for hybrid search and relationship-aware reranking.

## Prompt Augmentation in RAG

Apply query rewriting and instruction templating. Insert citations, require chain-of-thought internally, and use tool-call markers for structured tasks. For long contexts, use hierarchical chunk selection and budgeted summaries.

## Memory-Augmented Generation

Session memory stores user facts and preferences. Long-term memory summarizes history into canonical records. Use embeddings + recency decay to avoid stale or redundant memories.

## Evaluation Design

Evaluate with synthetic Q/A pairs plus human ratings for factuality and helpfulness. Automate with LLM judges, but sample audits are key. Track metrics like Context Recall, Faithfulness, Answer Quality, MRR@10, and nDCG@10.

## Latency Tuning

Use async fan-out for retrieval providers, cache frequent embeddings, and pre-warm vector indexes. Batch reranking requests and enforce token budgets per chunk.

## Security & Governance

Implement content safety filters, PII redaction, and audit logs. Support data residency and per-tenant indexes. Require signed prompts for sensitive tool calls, and maintain allow/deny lists for sources.

## Future Outlook

Lightweight retrievers, retrieval distillation, and synthetic corpus generation will reduce cost while increasing grounding reliability. Expect cross-modal retrieval and agentic multi-hop search to mature.