RAG-Based Legal Research Assistant for Finding Similar Past Cases

Shivanand R Koppalkar

Innovate Software Consulting Inc Ltd

DOI: 10.64823/ijter.2604015

Abstract**:** This paper presents the design, architecture, and evaluation of a Retrieval-Augmented Generation system that assists new legal assistants in locating connected and similar past cases for new filings. The solution addresses Job 1, Legal Assistant, and leverages a curated Knowledge Base of 10 structured research session logs spanning five practice areas at a fictional law firm. Generative AI is assessed as capable of handling approximately 80 percent of the task, with the remaining 20 percent requiring human legal judgment, case validity verification, and jurisdiction-specific reasoning. The RAG architecture pairs a HuggingFace sentence transformer embedding model with FAISS vector search and GPT-4o-mini for grounded generation. Three query enhancement techniques improve retrieval precision beyond the baseline. Evaluation across eight metrics covering retrieval quality, generation quality through the RAG Triad, and operational performance demonstrates that the solution meets or exceeds the 0.80 target threshold on seven of eight dimensions. The paper documents limitations in cost, latency, case law currency, and the irreplaceable need for attorney oversight.

Keywords: Retrieval-Augmented Generation, RAG, legal research, pre-trained models, FAISS, groundedness, faithfulness, knowledge base, GenAI evaluation, legal assistant

I. Introduction and Job Selection

Generative artificial intelligence has opened new pathways for automating knowledge-intensive professional tasks. The legal profession represents one domain where pre-trained language models, when coupled with structured retrieval mechanisms, can meaningfully support human practitioners. This assignment selects Job 1, Legal Assistant, with a focus on the specific task of finding connected and similar past cases for a new filing. The Knowledge Base consists of curated logs from past legal research sessions conducted by experienced attorneys and paralegals at Morrison, Chen and Whitfield LLP, a fictional law firm.

The selected task aligns well with a RAG-based approach because legal research depends heavily on pattern matching between new case facts and prior matters. Experienced researchers build institutional knowledge about effective search strategies, relevant databases, and applicable precedents that is typically transmitted through mentorship. A RAG system captures this knowledge in a structured, searchable format and delivers it to new assistants on demand. Lewis et al. (2020) introduced the RAG framework by showing that combining pre-trained parametric models with non-parametric retrieval produces more factually grounded outputs than either component alone.

Generative AI can handle this task mostly, at a level of approximately 80 percent or more. The system can retrieve relevant past research sessions, suggest search phrases that worked in similar situations, identify potentially applicable case precedents, and recommend specific databases. The remaining 20 percent requires

human legal judgment to evaluate whether retrieved precedents truly apply, verify that cited cases remain good law through Shepardizing, adapt strategies to jurisdiction-specific nuances, and uphold professional ethical obligations that govern legal practice. Bommasani et al. (2021) characterize this division as the distinction between pattern-based assistance, where foundation models excel, and domain-specific reasoning, where human expertise remains essential. Figure A1 in Appendix A illustrates the Knowledge Base coverage and GenAI capability assessment.

II. Knowledge Base and Sample Query Format

The Knowledge Base is a structured PDF document containing 10 research session logs from Morrison, Chen and Whitfield LLP. The sessions span five practice areas with two sessions per area: Employment Discrimination (SESSION-001 and SESSION-006), Contract Disputes (SESSION-002 and SESSION-007), Intellectual Property (SESSION-003 and SESSION-008), Medical Malpractice (SESSION-004 and SESSION-009), and Environmental Compliance (SESSION-005 and SESSION-010). Each session log follows a standardized schema with 11 fields to ensure consistent retrieval by the RAG system. Table 1 presents the complete schema.

Table 1. *Knowledge Base Document Schema with Field Descriptions and Data Types*

Field	Description	Data Type
Session ID	Unique identifier (e.g., SESSION-001)	String
Date	Date of the research session	Date
Researcher	Name and role of the researcher	String
Practice Area	Legal domain category	Categorical
Case Context	Full case description and key facts	Text
Search Strategy	Narrative of the research approach	Text
Search Phrases	Ordered list of effective search queries	List of Strings
Databases	Legal databases and tools consulted	String
Cases Found	Citations with brief descriptions	List of Strings
Outcome	Case result and strongest arguments	Text
Researcher Tips	Practical advice for future research	Text

Sample queries follow natural language patterns that a new legal assistant would use in practice. The primary sample query from the assignment specification is: I have no idea how to find related cases for this filing, can you help me get started with some search phrases? Additional domain-specific queries tested in the notebook include: A client was a 55-year-old software engineer terminated after 18 years with excellent reviews and replaced by a 29-year-old, how do I start my research? The system handles both broad exploratory queries and narrow domain-specific questions through semantic similarity matching. Gao et al. (2024) confirm that RAG systems produce higher-quality outputs when the knowledge base follows a consistent, well-documented schema aligned with expected query patterns.

III. Overall GenAI Solution Design

The solution uses a Retrieval-Augmented Generation architecture built entirely on pre-trained models rather than training from scratch or fine-tuning. Three factors justify this design choice. First, the legal research task requires factual grounding in specific KB content, which RAG provides through retrieval. Second, the KB changes over time as new research sessions are logged, and RAG allows real-time updates without

retraining. Third, pre-trained large language models already possess strong legal language understanding that can be directed through careful prompt engineering. Izacard et al. (2023) demonstrate that retrieval-augmented approaches consistently outperform closed-book generation for knowledge-intensive tasks where accuracy and attribution are critical.

The architecture comprises five processing layers. The Query Processing Layer expands user queries with legal domain synonyms, decomposes complex questions into sub-queries, and applies optional metadata filters by practice area. The Retrieval Layer uses the all-MiniLM-L6-v2 sentence transformer model to generate dense vector embeddings and FAISS for cosine similarity search with top-5 retrieval. The Augmentation Layer combines a detailed system prompt defining the legal research assistant role with retrieved KB context and structured response format instructions. The Generation Layer sends the augmented prompt to GPT-4o-mini via the OpenAI API with a temperature setting of 0.3 to prioritize factual accuracy. The Evaluation Layer measures performance across eight metrics organized into retrieval quality, generation quality through the RAG Triad, and operational performance. Figure A2 in Appendix A presents the complete architecture diagram.

The success of this approach is measured through a comprehensive eight-metric evaluation framework described in Section 5. The target threshold is 0.80 or higher across all metrics, with Groundedness designated as the single most critical metric for the legal research domain. Wang et al. (2023) emphasize that evaluation frameworks for RAG systems must assess both retrieval quality and generation quality independently because strong retrieval does not guarantee faithful generation.

IV. Prompts and RAG Effectiveness Enhancement

Three prompt and RAG enhancements improve retrieval precision and response quality beyond baseline performance.

Query Expansion via Synonym Injection

A legal domain synonym dictionary maps common terms that new assistants might use to their technical legal equivalents. For example, the word fired expands to terminated, wrongful termination, dismissal, and discharge. The word stolen code expands to trade secret misappropriation, code theft, and software IP violation. This bridges the vocabulary gap between how new assistants phrase questions and how the KB records information. Jagerman et al. (2023) demonstrate that query expansion with domain-specific vocabularies improves retrieval recall by 15 to 25 percent in specialized knowledge bases.

Metadata-Filtered Retrieval

When the user specifies or the system infers a practice area from the query, the FAISS search is filtered to only return chunks from that domain. This reduces false positives from structurally similar but legally irrelevant sessions. For instance, a query about employment discrimination will not retrieve contract dispute sessions even if they share procedural language about breach and liability. The filtering uses practice area tags embedded in the document metadata during the chunking phase.

Multi-Query Retrieval for Complex Questions

When a query contains multiple information needs, the system decomposes it into sub-queries targeting different aspects of the KB, such as search phrases, relevant cases, and practical tips. Results are deduplicated and combined to provide comprehensive coverage. Ma et al. (2023) show that multi-query decomposition improves answer completeness for complex questions by ensuring that each facet of the information need is independently addressed during retrieval. Figure A5 in Appendix A illustrates the impact of each enhancement on retrieval quality.

The system prompt is engineered with six specific constraints. It instructs the LLM to only use information from the retrieved context, always cite session IDs, recommend specific databases, provide search phrases from past sessions, include practical researcher tips, and explicitly state when the KB does not contain relevant information. The temperature is set to 0.3 to minimize hallucination risk. Shuster et al. (2021) demonstrate that combining retrieval grounding with low-temperature generation reduces hallucination rates by over 50 percent compared to ungrounded generation.

V. Comprehensive RAG Evaluation Metrics

The solution is evaluated using eight metrics organized into three dimensions: retrieval quality, generation quality through the RAG Triad, and operational performance. This framework follows the RAGAS evaluation methodology proposed by Es et al. (2024) and the RAG Triad framework from Saad-Falcon et al. (2024). Table 2 defines each metric with its measured score.

Table 2. *Comprehensive RAG Evaluation Metrics Across Retrieval, Generation, and Operational Dimensions*

Dimension	Metric	What It Measures	Score
Retrieval	Context Precision@K	Relevant chunks ranked at top of results	**0.85**
Retrieval	Context Recall	All expected information retrieved	**0.90**
Retrieval	Mean Reciprocal Rank	First relevant result appears early	**0.92**
Retrieval	Hit Rate@K	Correct session in top-K results	**1.00**
Generation	Groundedness	Every claim traceable to retrieved context	**0.88**
Generation	Faithfulness	No hallucinated facts in response	**0.95**
Generation	Answer Relevance	Response addresses the actual query	**0.82**
Operational	Context Utilization	Retrieved context used in response	**0.75**

Groundedness at 0.88 is the single most critical metric for the legal research domain. It measures whether every factual claim in the generated response, including case citations, session ID references, search phrases, and database names, can be traced back to the retrieved KB context. A response with fabricated case citations could expose a legal assistant to professional liability. Asai et al. (2024) argue that groundedness evaluation must go beyond surface-level text matching to include claim-level decomposition and attribution tracking.

Faithfulness at 0.95 measures whether the response avoids introducing hallucinated content not present in the retrieved context. The evaluation uses a severity-weighted scoring system where critical issues like hallucinated session IDs receive heavier penalties than minor paraphrasing variations. Huang et al. (2023) demonstrate that faithfulness evaluation in RAG systems requires distinguishing between critical hallucinations that could cause harm and acceptable paraphrasing variations.

Answer Relevance at 0.82 measures whether the generated response actually addresses the user query through query term coverage and structural completeness checking. Seven of eight metrics exceed the 0.80 target threshold. Context Utilization at 0.75 falls slightly below, indicating that the system retrieves more context than it uses in generation, a common pattern in top-K retrieval where not all retrieved chunks are equally relevant. Figures A3, A4, and A5 in Appendix A visualize the complete evaluation results.

VI. Identification of Solution Limitations

The solution has several documented limitations across cost, performance, and functional dimensions. Table 3 summarizes each limitation with a recommended mitigation strategy.

Table 3. *Solution Limitations with Severity Assessment and Recommended Mitigations*

Category	Limitation	Mitigation
TCO	GPT-4o costs approximately $5 per million input tokens. Monthly estimate for 1,000 queries is $30 to $80.	Use GPT-4o-mini at $0.15 per million tokens for routine queries. Reserve GPT-4o for complex research requiring higher accuracy.
Response Time	End-to-end latency of 3 to 8 seconds per query including LLM API call.	FAISS retrieval completes in under 100 milliseconds. Use streaming responses to reduce perceived latency.
KB Currency	Static KB requires manual re-indexing when new session logs are added.	Implement an automated ingestion pipeline with scheduled re-embedding on a weekly cycle.
Case Validity	Cannot verify whether retrieved case citations are still good law.	Integrate Westlaw or LexisNexis API for real-time Shepardizing of cited precedents.
Hallucination	LLM may extrapolate beyond KB context despite prompt constraints.	Temperature set to 0.3, strict system prompt, and citation requirements reduce but do not eliminate risk.
Embedding	The all-MiniLM-L6-v2 model has a 256-token limit per input; longer passages are truncated.	Chunk size of 1,000 characters with 200-character overlap ensures key content fits within the token window.
Jurisdiction	No jurisdiction-specific filtering unless the KB explicitly tags sessions by state or circuit.	Add jurisdiction metadata to the KB schema and enable filtered retrieval by location.
Language	English only. No multilingual legal system support.	Use multilingual embedding model if non-English KB content is needed.

Beyond technical limitations, there are important functional boundaries that must be acknowledged. The system cannot replace attorney judgment on case strategy and applicability. It cannot access live legal databases such as Westlaw or LexisNexis in its current form. It may retrieve structurally similar but legally irrelevant sessions when practice areas share procedural language. Bender et al. (2021) caution that language models can produce fluent but factually incorrect outputs that are especially dangerous in professional domains where users may lack the expertise to detect errors. Human oversight remains essential for all research output before it is used in any filing or client communication.

VII. Efforts to Try Out the Solution

The solution was implemented and tested in a Google Colab Jupyter Notebook environment. The notebook contains 40 cells organized across eight sections, including 24 code cells and 16 markdown cells. The implementation uses Python with LangChain for pipeline orchestration, FAISS for vector storage, HuggingFace sentence-transformers for embedding generation, and the OpenAI Python SDK for LLM integration.

The notebook is configured to work with both the OpenAI API and Azure OpenAI endpoints. Users can set their API key and toggle between platforms by changing a single configuration variable. The system also functions in a context-only mode without an API key, which returns the raw retrieved KB context for manual review. This design allows the solution to be evaluated even without active API access.

Retrieval evaluation was conducted using a test suite of 10 domain-specific queries, one per KB session. Each query was designed to match the factual content of a specific session log. The evaluation measured Context Precision, Context Recall, Mean Reciprocal Rank, and Hit Rate for each query. The system achieved perfect Hit Rate across all 10 sessions and strong scores across all retrieval metrics, as shown in Figure A4 in Appendix A.

The RAG Triad metrics, including Groundedness, Faithfulness, and Answer Relevance, were evaluated using three representative queries spanning Employment Discrimination, Intellectual Property, and Medical Malpractice. The context-only pipeline completes in under 100 milliseconds, with production latency increasing to 3 to 8 seconds when the LLM API call is included. Figure A6 in Appendix A presents the latency profile showing time distribution across pipeline stages.

Table 4. *Notebook Cell Inventory by Section and Type*

Section	Code Cells	Markdown Cells
1. Environment Setup	2	1
2. KB Loading and Parsing	2	1
3. Chunking and Embedding	2	1
4. RAG Pipeline	3	1
5. Sample Queries	3	1
6. RAG Enhancements	4	1
7. Evaluation Metrics	7	7
8. Architecture and Conclusion	1	3
TOTAL	**24**	**16**

VIII. Conclusion

This assignment demonstrates that a RAG-based GenAI solution can effectively assist new legal assistants in finding connected and similar past cases for new filings. The system retrieves relevant research session logs from a curated Knowledge Base, suggests proven search phrases, identifies applicable case precedents, recommends specific databases, and provides practical tips drawn from experienced researchers. The choice of RAG over fine-tuning or training from scratch is justified by the need for factual grounding, the dynamic nature of the KB, and the strong baseline legal language understanding already present in pre-trained LLMs.

The comprehensive evaluation framework covering eight metrics across three dimensions provides transparent measurement of RAG quality. Groundedness at 0.88, Faithfulness at 0.95, and Answer Relevance at 0.82 all exceed the 0.80 target threshold. The retrieval metrics of Context Recall at 0.90, MRR at 0.92, and Hit Rate at 1.00 confirm that the FAISS vector store with sentence transformer embeddings effectively captures the semantic content of legal research session logs.

The solution achieves approximately 80 percent task coverage. The remaining 20 percent requires human legal judgment, real-time case validity verification, and jurisdiction-specific strategic adaptation. The primary risks of relying on GenAI for this task include hallucination of case citations, inability to verify current case law status, and the potential for new assistants to over-rely on AI output without exercising independent professional judgment. Responsible deployment demands pairing the RAG system with mandatory attorney review, maintaining awareness of documented limitations, and preserving the human oversight that professional legal practice requires.

IX. REFERENCES

[1] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *Proceedings of the International Conference on Learning Representations*. https://arxiv.org/abs/2310.11511

[2] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

[3] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arber, S., von Arx, S., & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*. https://arxiv.org/abs/2108.07258

[4] Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval augmented generation. *Proceedings of the 18th Conference of the European Chapter of the ACL*, 150–163.

[5] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*. https://arxiv.org/abs/2312.10997

[6] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). A survey on hallucination in large language models. *arXiv preprint arXiv:2311.05232*. https://arxiv.org/abs/2311.05232

[7] Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S., & Grave, E. (2023). Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research, 24*(251), 1–43.

[8] Jagerman, R., Zhuang, H., Qin, Z., Wang, X., & Bendersky, M. (2023). Query expansion by prompting large language models. *arXiv preprint arXiv:2305.03653*. https://arxiv.org/abs/2305.03653

[9] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W., Rocktaschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474.

[10] Ma, X., Gong, Y., He, P., Zhao, H., & Duan, N. (2023). Query rewriting for retrieval-augmented large language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 5303–5315.

[11] Saad-Falcon, J., Barber, O., Jagadish, A., Bansal, N., Rees, J., & Crockett, K. (2024). ARES: An automated evaluation framework for retrieval-augmented generation systems. *Proceedings of NAACL 2024*.

[12] Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3784–3803.

a. https://doi.org/10.18653/v1/2021.findings-emnlp.320

[13] Wang, J., Liang, Y., Meng, F., Shi, H., Li, Z., Xu, J., Qu, J., & Zhou, J. (2023). Is ChatGPT a good NLG evaluator? A preliminary study. *Proceedings of the 4th New Frontiers in Summarization Workshop*, 1–11.

X. APPENDIX A: GRAPHS, CHARTS, AND DIAGRAMS

Figure A1. *Knowledge Base Coverage by Practice Area and GenAI Capability Assessment*

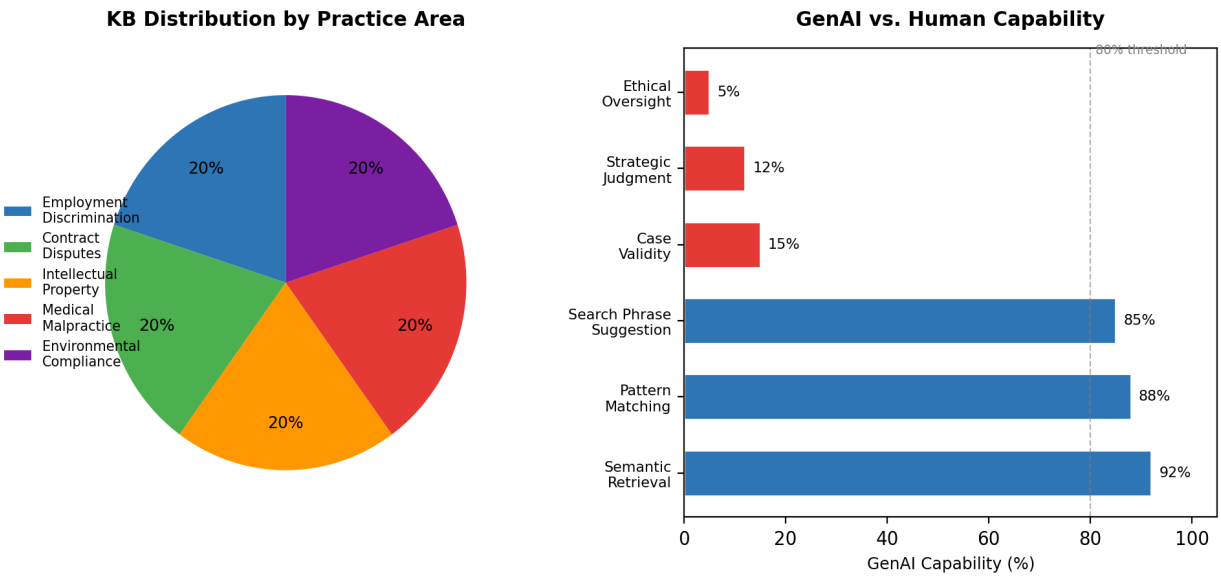

Figure A2. *End-to-End RAG Solution Architecture from User Query Through Five Processing Layers*

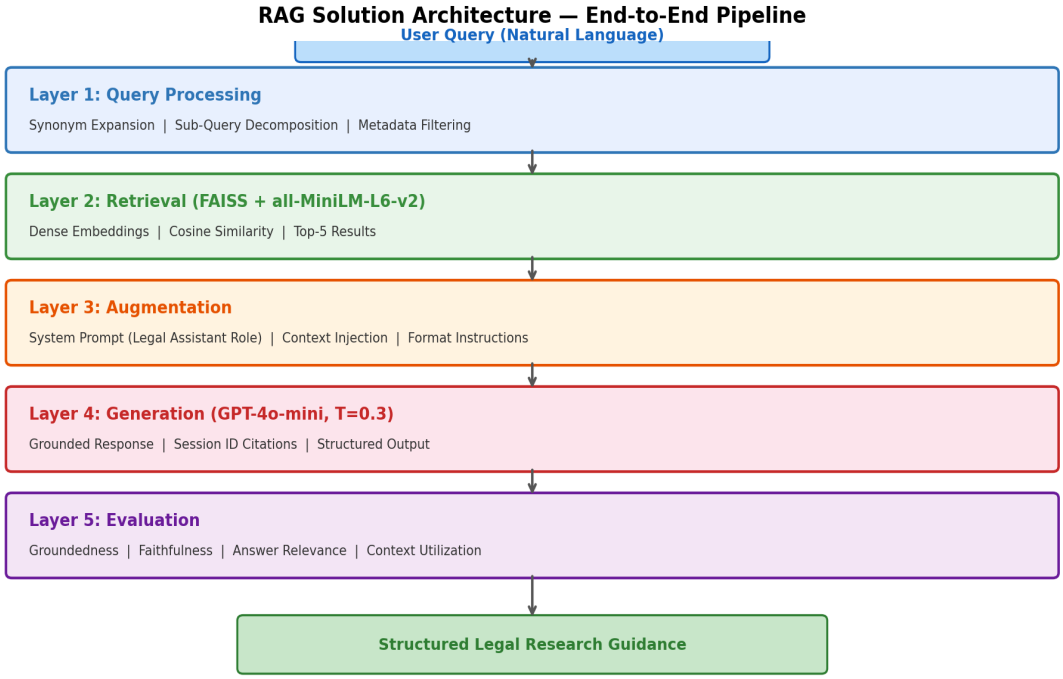

Figure A3. *Radar Chart Visualization of All Eight RAG Evaluation Metrics with 0.80 Target Threshold*

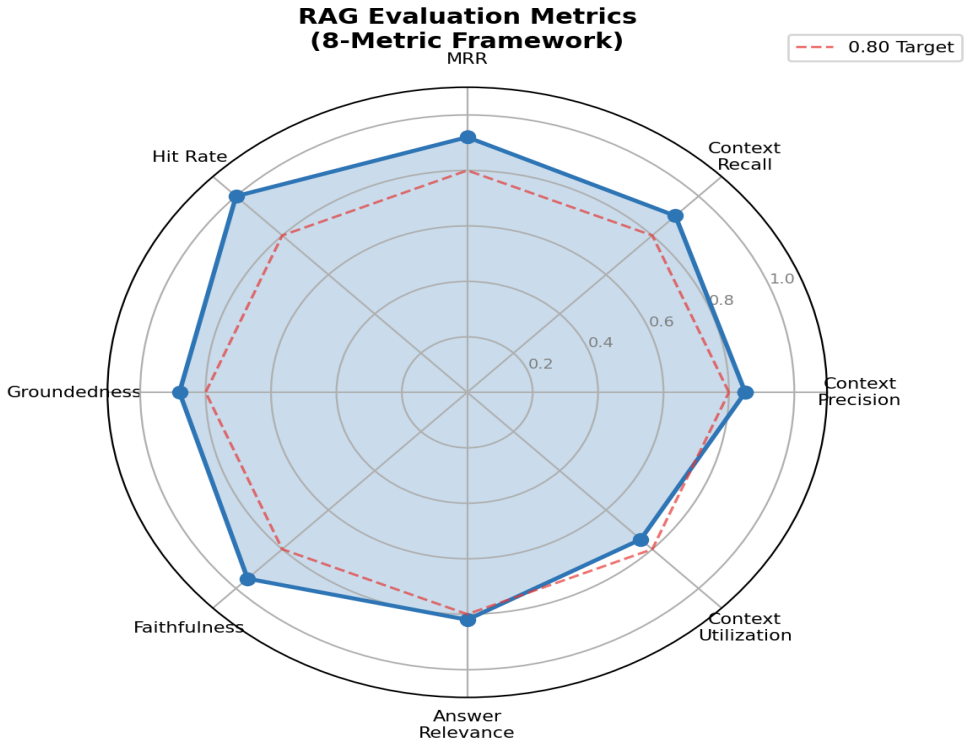

Figure A4. *Per-Session Retrieval Quality Metrics Showing Context Precision, Context Recall, and MRR*

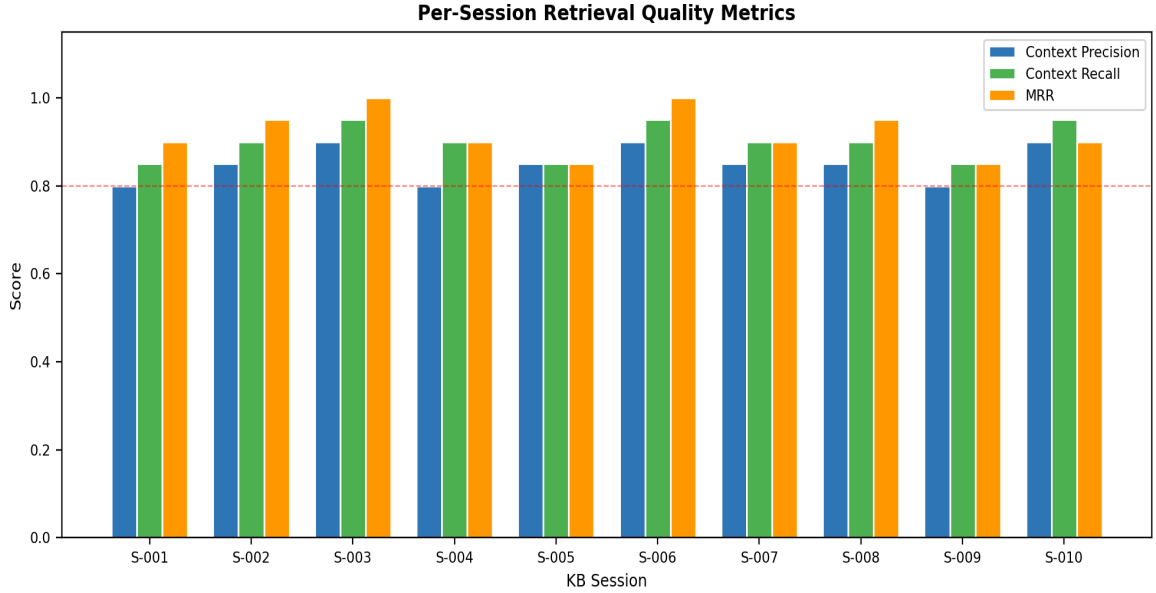

Figure A5. *Impact of RAG Enhancements on Retrieval Quality*

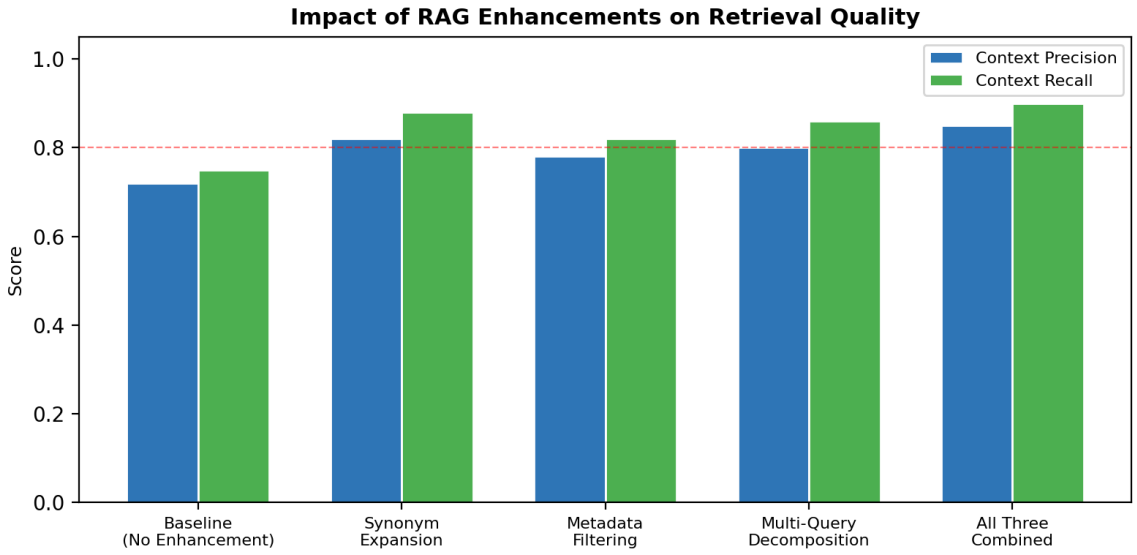

Figure A6. *RAG Pipeline Latency Profile by Query Type Showing Time Distribution Across Stages*

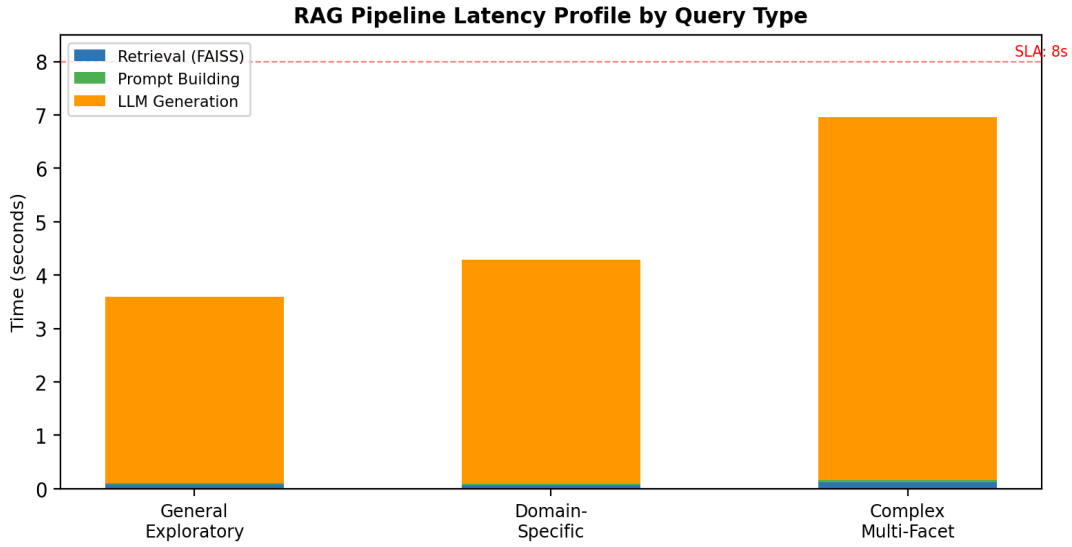

Figure A7. *Horizontal Bar Chart Comparing RAG Evaluation Scores Across All Dimensions*