Table of contents

1. Introduction

Large Language Models (LLMs) excel at natural language tasks but are limited by static pretraining data, lacking access to current/domain-specific knowledge. Retrieval-Augmented Generation (RAG) addresses this by integrating external knowledge; it retrieves relevant documents and incorporates them into prompts, enabling grounded, accurate, and verifiable responses (Karpukhin et al., 2020).

This project develops an advanced RAG pipeline incorporating three core enhancements: semantic chunking, hybrid retrieval, and cross-encoder reranking. Unlike traditional character-based chunking, semantic chunking preserves sentence boundaries, improving retrieval precision and generation quality (Asghar et al., 2024).

Hybrid retrieval combines dense vector similarity with sparse keyword-based methods to capture both conceptual meaning and exact legal phrasing (Sharma et al., 2024). Cross-encoder reranking further refines top retrieved passages by encoding queries and candidates together, creating more accurate relevance scoring (Sanseviero, 2024).

2. Domain selection & dataset

This project leverages the General Data Protection Regulation (GDPR). As a formally structured, linguistically dense, legislative text with rich internal referencing, GDPR is ideal for testing retrieval precision and generation reliability, requiring high accuracy in interpretation (Ariai and Demartini, 2025).

Legal queries frequently require multi-source insights, ideal for assessing multi-hop retrieval and reranking. Since LLMs offer only partial or outdated legal knowledge from pretraining, RAG's ability to produce grounded, verifiable outputs from authoritative documents becomes crucial (Monroy et al., 2024).

The corpus was compiled from three complementary sources: the UK GDPR (primary legislative text), UK Information Commissioner’s Office (ICO) Guidance (official interpretation), and CNIL GDPR Practice Guide (technical implementation). Collectively, these documents provide regulatory, interpretive, and operational facets of GDPR, enabling accurate responses to diverse legal queries. While general LLMs can approximate answers, a RAG system ensures traceability to precise legal texts and interpretations, which is critical for auditability and accuracy in law.

3. Pipeline architecture

The RAG pipeline consists of six key stages to enhance retrieval and generation quality in a legal domain QA setting.

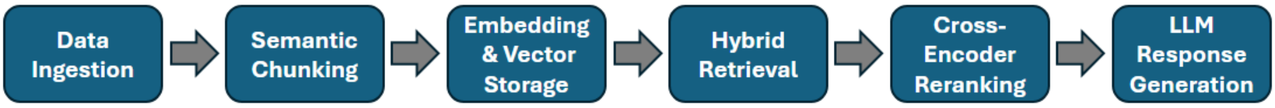


Figure 1: RAG pipeline architecture

1. **Data Ingestion:** The three legal documents are ingested via upload into Colab.

2. **Semantic Chunking:** Documents are split into sentence-level chunks using **NLTKTextSplitter** to ensure coherent and context-preserving segments for embedding.

3. **Embedding & Vector Storage:** Chunks are encoded using **BAAI/bge-base-en** into 768-dimensional vectors and stored in a persistent Chroma vector database for fast similarity search.

4. **Hybrid Retrieval:** Two parallel retrievers are used: **BM25** for exact keyword matching and **vector similarity search** for semantic relevance. Results are merged and deduplicated.

5. **Cross-Encoder Reranking:** Top retrieved candidates are re-ranked using a cross-encoder (**bge-reranker-large**) to prioritize passages with the highest contextual relevance.

6. **LLM Response Generation:** The top-ranked chunks are passed to **Gemini Flash 2.0**, generating answers grounded strictly in the retrieved context.

4. Design justification & methodological choices

This section outlines the key design decisions across the RAG pipeline to maximize retrieval accuracy, semantic coherence, and factual correctness – core requirements in domains like legal compliance.

4.1 Semantic chunking

To preserve logical meaning in legal text, sentence-aware chunking was applied using **NLTK's TextSplitter**. This approach avoids cutting clauses mid-sentence, maintaining legal coherence. Chunks averaged 800 tokens with a 100-token overlap, and occasional overflows were tolerated to preserve sentence coherence. This improved both retrievability and semantic precision.

4.2 Embedding model & vector storage

Dense embeddings were generated using **BAAI/bge-base-en**, a 768-dimensional model optimized for English semantic search. Compared to smaller models like MiniLM, BGE improves zero-shot retrieval – particularly useful for legally nuanced queries. Embeddings were stored in ChromaDB, a high-performance vector database suitable for legal-scale workloads.

4.3 Hybrid retrieval

A hybrid strategy was adopted to capture both linguistical precision and semantic depth. **Dense retrieval** captured paraphrased or conceptually related content, while **BM25** ensured exact term matching. Results from both methods were merged and deduplicated. This combination is well-supported for legal QA tasks, where both recall and conceptual coverage matter (Sharma et al., 2024).

4.4 Cross-encoder reranking

Retrieved results were reranked using **BAAI/bge-reranker-large**, a cross-encoder that processes the query and each document together to assess contextual fit. Unlike bi-encoders, this method captures fine-grained semantic interactions, improving precision – especially for complex or ambiguous legal queries (Sanseviero, 2024).

4.5 LLM response generation

Final responses were generated using Gemini Flash 2.0, selected for its inference speed and consistent handling of structured prompts. The prompt template instructed the model to rely solely on the provided context and refuse to answer when evidence was insufficient. With a temperature of 0.3 and a 1024-token cap, outputs remained focused, grounded, and auditable – crucial in legal applications.

5. Evaluation & testing

5.1 Evaluation approach

Evaluating a RAG system involves assessing both retrieval accuracy and generation quality, making it more complex than standard LLM evaluation (Lewis et al., 2020; Wu et al., 2022). This evaluation framework consists of:

- **Retrieval Evaluation:** Measures whether retrieved passages contain necessary information.
- **Generation Evaluation:** Assesses factual correctness, source faithfulness, and response completeness.
- **Human Evaluation:** Judges the overall usefulness, clarity, and trustworthiness of the output.

5.2 Evaluation design

A set of 12 queries at three difficulty levels was developed to test a range of query types – from factual to nuanced and interpretive. Given the legal context and lack of labeled ground truth, qualitative, manual evaluation was used, following best practices in compliance-oriented NLP research (Miao et al., 2025).

- **Retrieval Relevance:** Did the documents match the query intent?
- **Faithfulness:** Was the output grounded in the retrieved context?
- **Completeness:** Did the response fully address the query?

Each dimension was scored on a 0-5 scale. Aggregated scores across all 12 queries reflect the model's overall performance. See Appendix 2 (Difficulty level definitions), Appendix 3 (Scoring definitions), Appendix 4 (Query set), and Appendix 5 (Model variants, Figures 5.1 - 5.4) for full breakdowns.

6. Results

As shown in Figure 2, model performance improved consistently with each variant augmentation. The simplest version (Baseline A), using only dense retrieval with MiniLM, scored 82 points. The most advanced configuration (Final Model), which integrated semantic chunking, hybrid retrieval, BGE embeddings, and cross-encoder reranking, scored 128 – a 56% improvement.

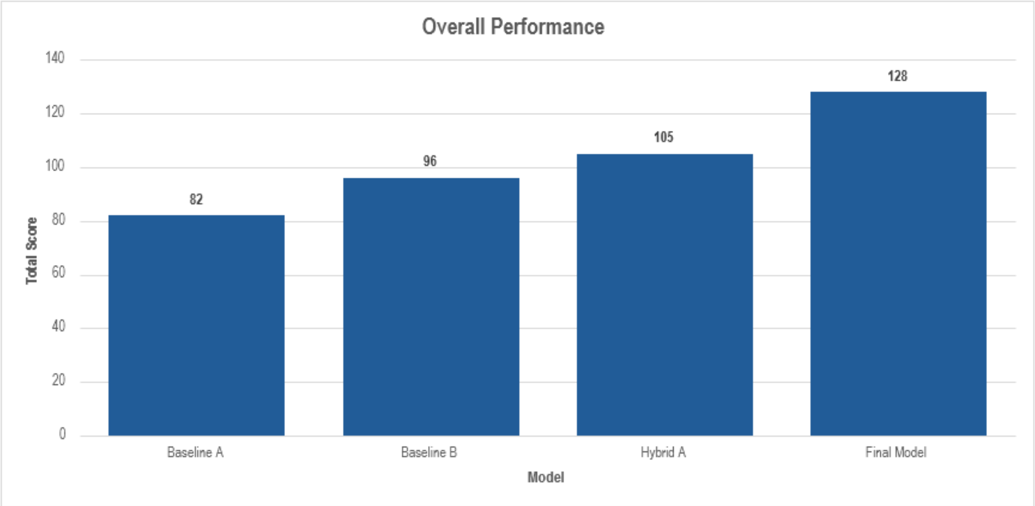


Figure 2. Total scores across all model variants on 12 GDPR queries.

Across all models (Figure 3), retrieval relevance consistently received the highest scores, confirming that each setup could retrieve broadly relevant content. However, faithfulness and completeness saw only modest improvements between Baseline B (vector + reranker, 96 points) and Hybrid A (BM25 + vector search, 105 points). This indicates that reranking and hybrid retrieval separately improved the model slightly, but only when combined in the Final Model did they significantly improve the model.

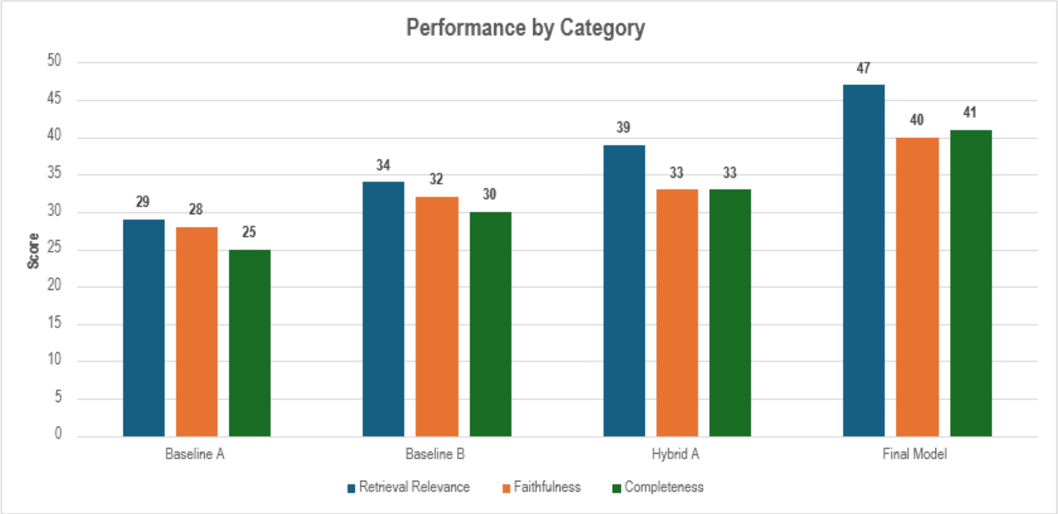


Figure 3. Scores by category across all model variants on 12 GDPR queries.

As illustrated in Figure 4, performance declined slightly with increased question difficulty, averaging a 1-point drop from Easy to Hard. This reflects the expected growing interpretive demands of complex legal queries. Importantly, no model showed severe degradation, and the Final Model maintained strong results even on the hardest prompts (average score: 10.0/15).

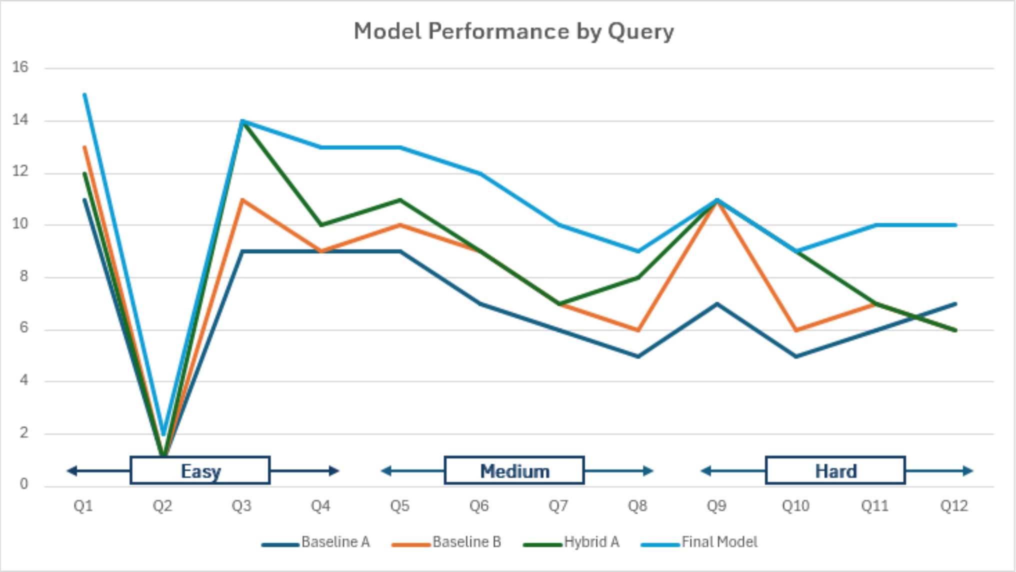


Figure 4. Scores by query across all model variants on 12 GDPR queries.

A consistent failure was Q2, which none of the models answered successfully. This was due to the lack of an explicit definition of “purpose” in the corpus and the system’s strict instruction against hallucination. While this preserves legal reliability, it highlights a core limitation: RAG systems may underperform on queries requiring general background knowledge not present in source documents.

7. Limitations & future work

While this project demonstrates the strengths of RAG in legal QA, several limitations highlight areas for future development.

7.1 System performance & scalability

Evaluation was limited by the system's inability to process multiple queries in batches, with each taking around one minute. For production use, especially with complex or chained queries, faster and more scalable solutions are required.

7.2 Corpus size & real-world complexity

The prototype operated on a 465-page corpus. Real-world legal applications often involve thousands of documents, demanding more advanced chunking, retrieval, and indexing, as well as stronger computational resources and commercial-grade models.

7.3 Limitations in abstract or Implicit queries

All models failed to answer Q2, as the corpus lacked an explicit definition and hallucination was strictly prohibited. While this reinforces legal reliability, it exposes a weakness when general knowledge is missing. Future systems should explore fallback strategies (i.e. predefined responses) for foundational queries.

7.4 Evaluation scope and realism

Only 12 queries and a limited number of model variants were evaluated. A more robust assessment would involve larger, more diverse question sets, expert annotation, user testing, and iterative feedback – all of which were beyond this project’s scope.

7.5 Baseline LLM performance & comparison

Testing with Gemini Flash 2.5 and GPT-4 showed strong performance without retrieval. These models likely benefit from pretraining on GDPR but lack verifiability and citation. In legal contexts, this remains a key limitation. Future research could explore hybrid systems that combine fluent generation with document-grounded verification.

7.6 Future directions

Future iterations should consider:

- Confidence scoring for uncertain outputs.
- Metadata-aware hybrid retrieval.
- Controlled fallback to general knowledge.
- Expanded tasks (i.e. summarization, clause comparison, drafting).

While large LLMs excel at general legal Q&A, RAG remains essential where traceability, document-specific reasoning, and factual reliability are critical – especially in domains with proprietary or rapidly evolving content.

8. References

Ariai, F. and Demartini, G. (2025) *Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges.* arXiv preprint arXiv:2410.21306. Available at: http://arxiv.org/pdf/2410.21306 (Accessed: 1 June 2025).

Asghar, M., Siddiqui, A.A., Khan, F.A. and Tahir, A. (2024) *Enhancing RAG Performance Through Chunking and Text Splitting Techniques.* Available at: https://www.researchgate.net/publication/383998204_Enhancing_RAG_Performance_Through_Chunking_and_Text_Splitting_Techniques (Accessed: 1 June 2025).

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wenzek, G., Lachaux, M.A., Wu, B., Edunov, S., Grave, E., Poesia, J. and Yih, W.T. (2020) *Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6718–6734.

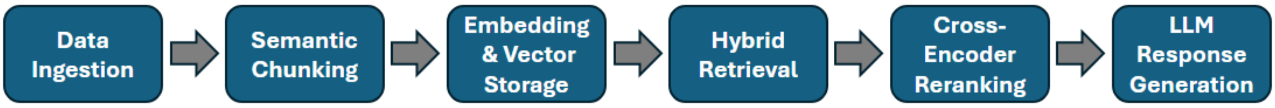
Miao, G., Chen, B., Wang, X., Xu, X., Zhang, K., Zhang, H., Lu, H., Wu, Y., Liu, P., Yang, D., Feng, J. and Wen, R. (2025) *Retrieval Augmented Generation Evaluation in the Era of Large Language Models: A Comprehensive Survey.* arXiv preprint arXiv:2504.14891. Available at: https://arxiv.org/html/2504.14891v1 (Accessed: 3 June 2025).

Monroy, M., Lupu, M., Bozzon, A. and Alam, M. (2024) *Measuring the Groundedness of Legal Question-Answering Systems.* arXiv preprint arXiv:2410.08764v1. Available at: https://arxiv.org/html/2410.08764v1 (Accessed: 4 June 2025).

Sanseviero, O. (2024) *Sentence Embeddings. Cross-encoders and Re-ranking.* [Blog] hackerllama - GitHub Pages. Available at: https://osanseviero.github.io/hackerllama/blog/posts/sentence_embeddings2/ (Accessed: 3 June 2025).

Sharma, A., Gupta, R., Jain, A. and Singh, S. (2024) *Domain-specific Question Answering with Hybrid Search.* Available at: https://arxiv.org/html/2412.03736v1 (Accessed: 3 June 2025).

9. Appendices



Appendix 1 - Figure 1: RAG pipeline architecture.

Source: own creation

Category: Easy (Direct Factual Recall)

These queries test the RAG's ability to retrieve and present straightforward definitions or core concepts explicitly stated in the corpus.

Category: Medium (Interpretive / Compliance / Conceptual)

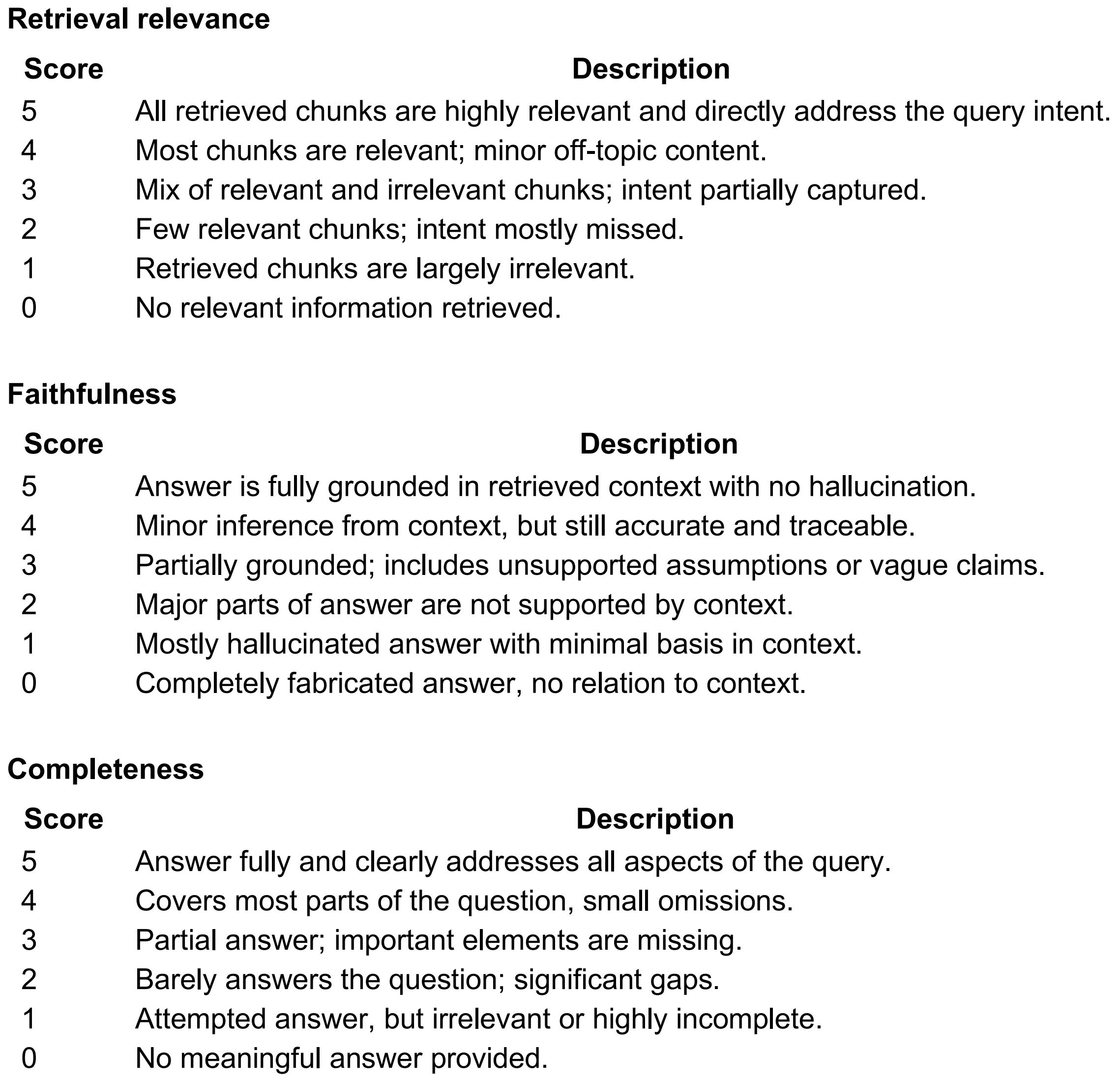
These queries require understanding of implications, specific compliance procedures, or distinguishing between closely related concepts. They go beyond simple definitions but don't involve complex multi-article synthesis or ambiguous scenarios.

Category: Hard (Application / Nuanced Legal Interpretation / Scenario-based)

These queries demand deeper understanding, synthesis of information from multiple parts of the law/guides, application of principles to hypothetical (or complex real-world) scenarios, or navigating potential ambiguities and edge cases. These are designed to truly stress-test the RAG's ability to handle domain-specific complexity.

Appendix 2: Difficulty level definitions

Source: own creation



Retrieval relevance

Score	Description
5	All retrieved chunks are highly relevant and directly address the query intent.
4	Most chunks are relevant; minor off-topic content.
3	Mix of relevant and irrelevant chunks; intent partially captured.
2	Few relevant chunks; intent mostly missed.
1	Retrieved chunks are largely irrelevant.
0	No relevant information retrieved.

Faithfulness

Score	Description
5	Answer is fully grounded in retrieved context with no hallucination.
4	Minor inference from context, but still accurate and traceable.
3	Partially grounded; includes unsupported assumptions or vague claims.
2	Major parts of answer are not supported by context.
1	Mostly hallucinated answer with minimal basis in context.
0	Completely fabricated answer, no relation to context.

Completeness

Score	Description
5	Answer fully and clearly addresses all aspects of the query.
4	Covers most parts of the question, small omissions.
3	Partial answer; important elements are missing.
2	Barely answers the question; significant gaps.
1	Attempted answer, but irrelevant or highly incomplete.
0	No meaningful answer provided.

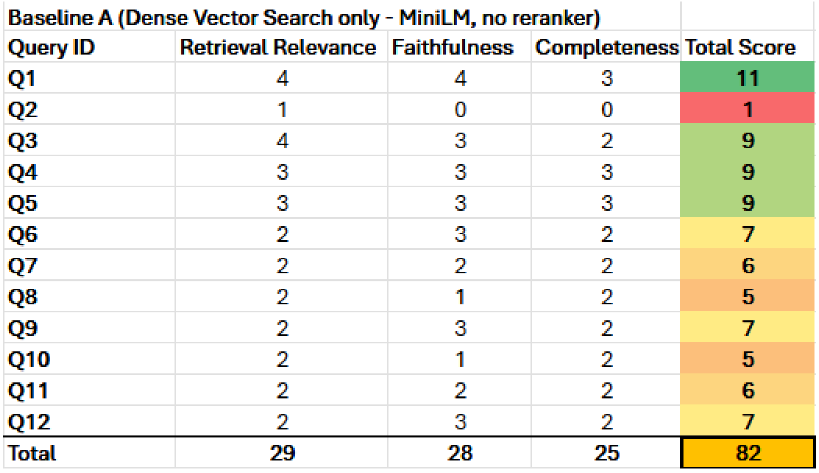
Appendix 3: Scoring definitions

Source: own creation

Query ID	Query	Difficulty	Type
Q1	What are the lawful bases for processing personal data under GDPR?	Easy	Direct Factual
Q2	What is the purpose of the GDPR?	Easy	Direct Factual
Q3	What does the GDPR say about consent?	Easy	Direct Factual
Q4	What is personal data according to the GDPR?	Easy	Direct Factual
Q5	When is a Data Protection Impact Assessment (DPIA) required under the GDPR, and what are its key components?	Medium	Compliance / Procedural
Q6	Explain the principle of 'storage limitation' and provide an example of how an organization might implement it.	Medium	Interpretive / Conceptual
Q7	What are the roles and responsibilities of a Data Protection Officer (DPO) under the GDPR?	Medium	Compliance / Role-based
Q8	How does the right to object to processing differ from the right to restriction of processing under GDPR?	Medium	Comparative / Conceptual
Q9	When transferring personal data from the UK to a country without an adequacy decision, what are the primary appropriate safeguards available under GDPR, and what essential assessment must a controller conduct regarding the recipient country's laws before relying on such safeguards?	Hard	Application / Cross-border Transfers
Q10	Describe the specific conditions under which automated individual decision-making, including profiling, is permissible under GDPR, and outline the key rights individuals have in relation to such decisions, especially when they produce legal effects or similarly significant effects.	Hard	Application / Rights & Conditions
Q11	Explain the definitions of a data controller, a data processor, and a joint controller under GDPR. What are the key differentiating responsibilities of each role, particularly concerning accountability and the lawful basis for processing?	Hard	Nuanced Legal Interpretation
Q12	Beyond administrative fines, list and briefly describe the other significant corrective powers that a supervisory authority (like the ICO) can impose for non-compliance with GDPR, including measures related to processing operations and data subject rights.	Hard	Compliance / Comprehensive Sanctions

Appendix 4: Query set

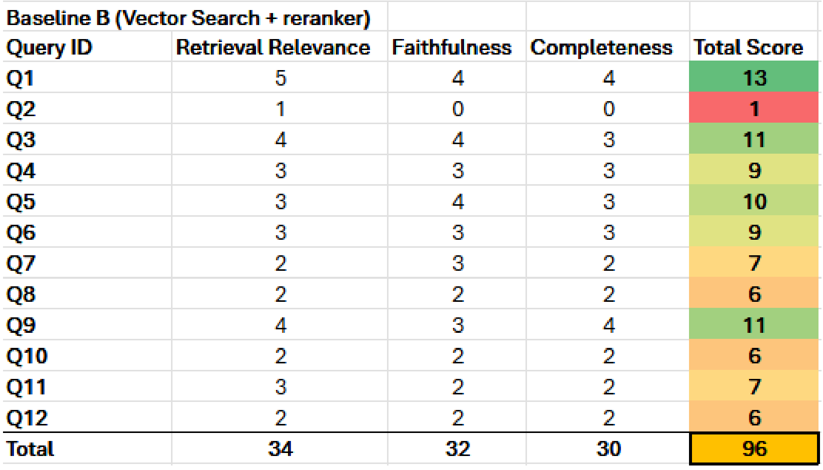
Source: own creation



Baseline A (Dense Vector Search only - MiniLM, no reranker)

Query ID	Retrieval Relevance	Faithfulness	Completeness	Total Score
Q1	4	4	3	11
Q2	1	0	0	1
Q3	4	3	2	9
Q4	3	3	3	9
Q5	3	3	3	9
Q6	2	3	2	7
Q7	2	2	2	6
Q8	2	1	2	5
Q9	2	3	2	7
Q10	2	1	2	5
Q11	2	2	2	6
Q12	2	3	2	7
Total	29	28	25	82

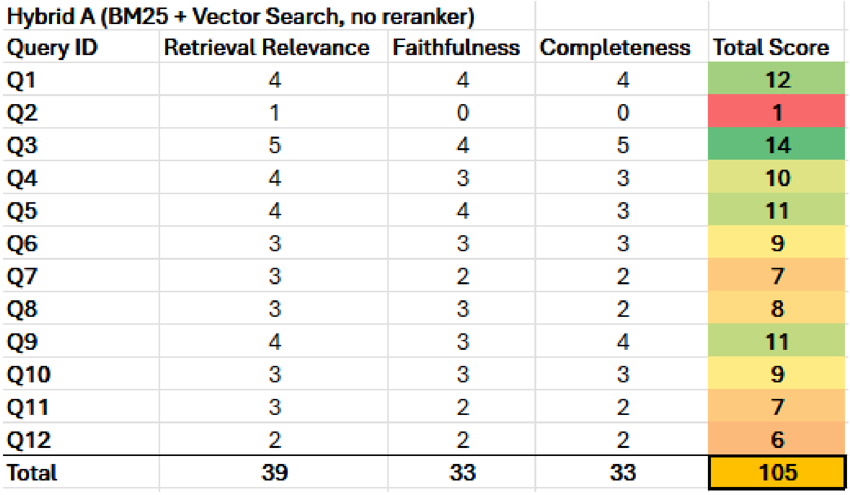
Figure 5.1: Baseline A evaluation table



Baseline B (Vector Search + reranker)

Query ID	Retrieval Relevance	Faithfulness	Completeness	Total Score
Q1	5	4	4	13
Q2	1	0	0	1
Q3	4	4	3	11
Q4	3	3	3	9
Q5	3	4	3	10
Q6	3	3	3	9
Q7	2	3	2	7
Q8	2	2	2	6
Q9	4	3	4	11
Q10	2	2	2	6
Q11	3	2	2	7
Q12	2	2	2	6
Total	34	32	30	96

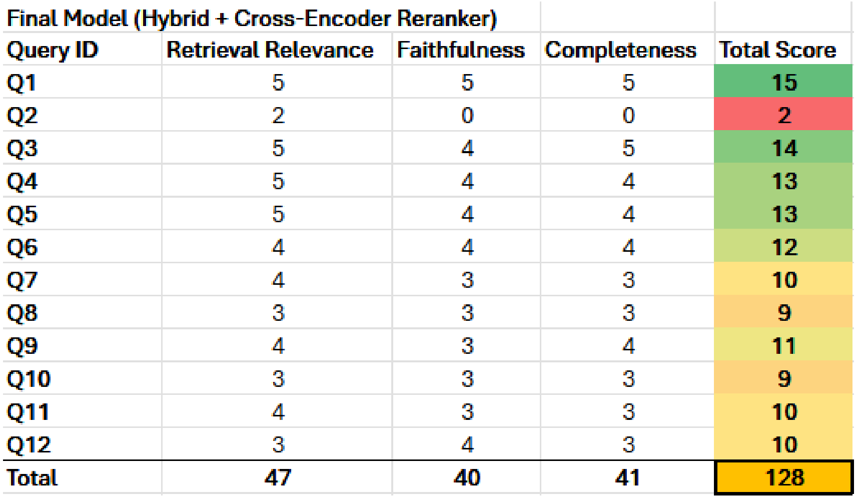
Figure 5.2: Baseline B evaluation table



Hybrid A (BM25 + Vector Search, no reranker)

Query ID	Retrieval Relevance	Faithfulness	Completeness	Total Score
Q1	4	4	4	12
Q2	1	0	0	1
Q3	5	4	5	14
Q4	4	3	3	10
Q5	4	4	3	11
Q6	3	3	3	9
Q7	3	2	2	7
Q8	3	3	2	8
Q9	4	3	4	11
Q10	3	3	3	9
Q11	3	2	2	7
Q12	2	2	2	6
Total	39	33	33	105

Figure 5.3: Hybrid A evaluation table

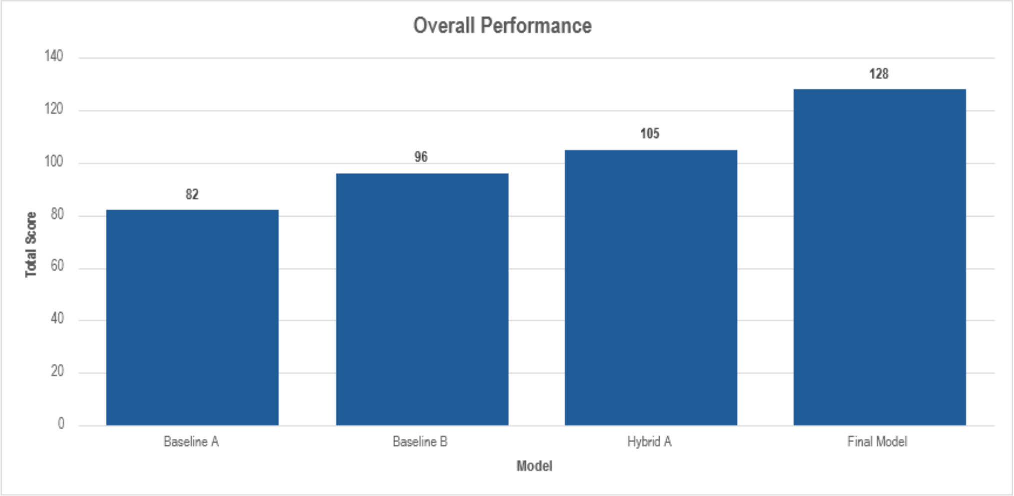


Final Model (Hybrid + Cross-Encoder Reranker)				
Query ID	Retrieval Relevance	Faithfulness	Completeness	Total Score
Q1	5	5	5	15
Q2	2	0	0	2
Q3	5	4	5	14
Q4	5	4	4	13
Q5	5	4	4	13
Q6	4	4	4	12
Q7	4	3	3	10
Q8	3	3	3	9
Q9	4	3	4	11
Q10	3	3	3	9
Q11	4	3	3	10
Q12	3	4	3	10
Total	47	40	41	128

Figure 5.4: Final model evaluation table

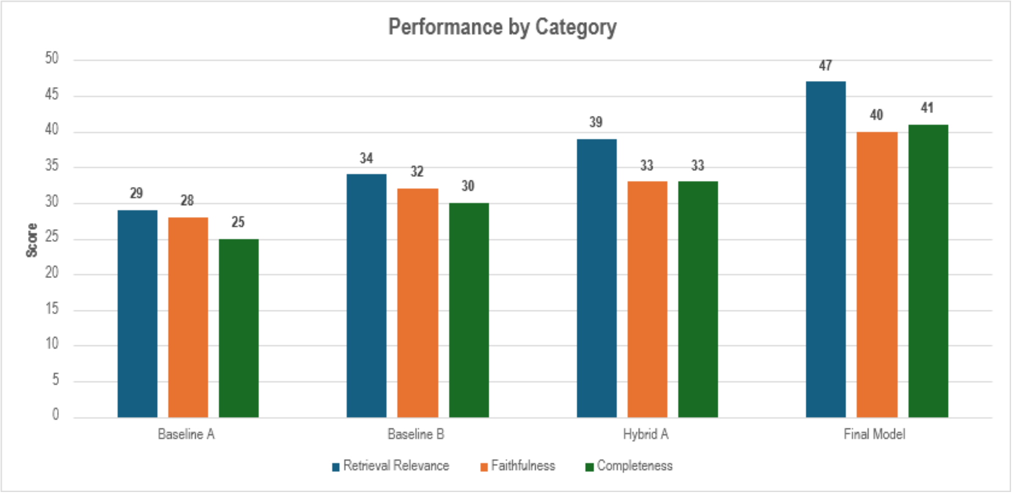
Appendix 5: Model variations

Source: own creation



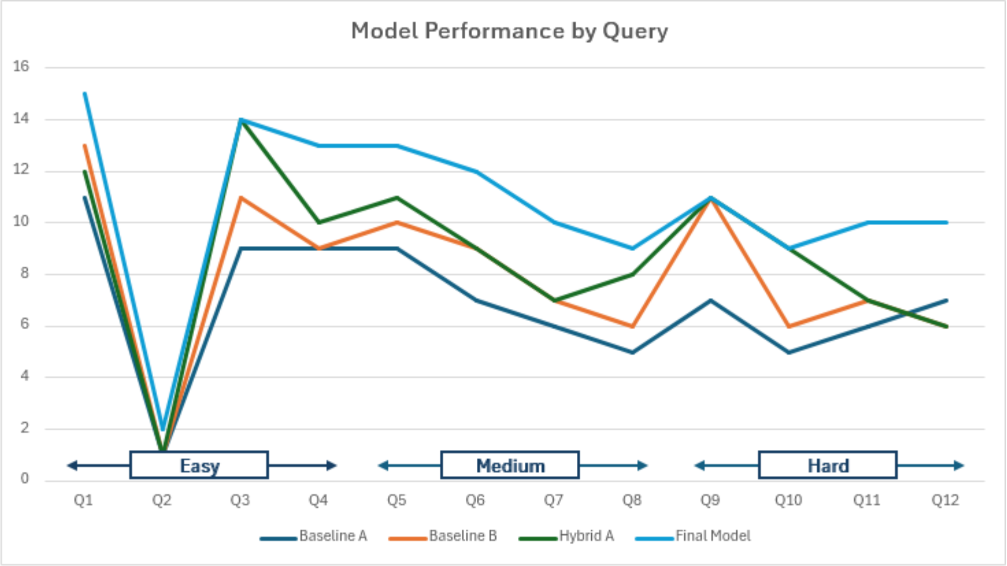
Appendix 6 – Figure 2. Total model scores.

Source: own creation



Appendix 7 – Figure 3. Scores by category.

Source: own creation



Appendix 8 – Figure 4. Scores by queries.

Source: own creation