Retrieval-augmented Generation for Academic Research

Milestone 2 Report

Group: QueryMinds
Students: Mohammadali Dehghani, Amir Saadati, Amina Kadic, Meliha Kasapovic

November 25, 2025

1 Introduction

The goal of Milestone 2 is to implement and evaluate several baseline methods for the task of ranking relevant text chunks relative to a given question. All methods return a ranking of the most relevant segments and use a common set of questions, text chunks, and manually annotated gold labels.

Four groups of baseline models have been developed:

- **Keyword Overlap** – simple keyword overlap counting,
- **TF–IDF Cosine** – vectorization and cosine similarity,
- **Embedding model** – SentenceTransformers MiniLM-L6-v2,
- **Supervised Classifier** – TF–IDF + logistic regression (balanced classes).

2 Data and Processing

The dataset consists of scientific PDF articles segmented into shorter text units. (200 tokens per chunk with 50-token overlap). From the 30 selected papers, this process produced approximately 450 chunks. Each chunk has its own ID and text, while questions contain text, optional keywords, and a unique identifier. Gold annotations are binary values $(q\_id, c\_id, label)$ that indicate relevance. The evaluation uses a curated set of 15 questions, provided in `data/questions.json`.

3 Methods

3.1 Rule-Based Approaches

The keyword method uses a simple token overlap. The TF–IDF approach measures the cosine similarity between questions and chunks. These methods are fast, but limited in semantics.

3.2 Embedding baseline

The `all-MiniLM-L6-v2` model is used for semantic similarity. Questions and chunks are encoded into vectors, and then compared using the cosine measure.

3.3 Supervised Model

A binary classifier is built based on:

- TF–IDF representation of the merged text (question + chunk),
- logistic regression with balanced classes.

The model returns the probability of relevance, based on which a ranking list is formed. The dataset is split into 80% training and 20% testing, using a stratified split to preserve label proportions. Logistic regression is trained with `class_weight='balanced'` to address class imbalance.

The model returns the probability of relevance, based on which a ranking list is formed.

4 Quantitative Evaluation

Evaluation metrics include: Precision, Recall, F1, Accuracy, and Precision@5 and Recall@5.

4.1 Results table

Baseline	**Prec**	**Rec**	**F1**	**Acc**	**P@5**	**R@5**
keyword overlap	0.542	0.840	0.613	0.547	0.520	0.277
tfidf cosine	0.533	0.857	0.602	0.561	0.613	0.347
embedding	0.475	1.000	0.609	0.479	0.573	0.339
supervised classifier	0.473	1.000	0.607	0.473	0.747	0.497

Table 1: Comparative performance table of all baseline methods.

4.2 Visualizations

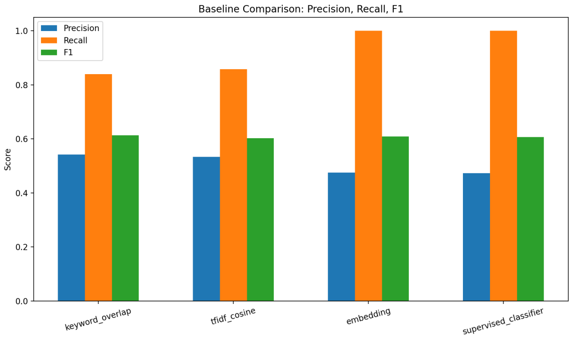

Figure 1: Precision, Recall and F1.

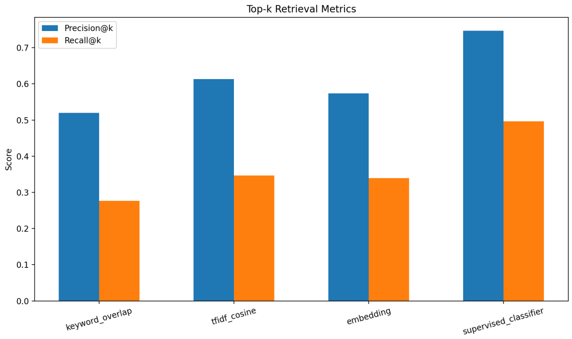

Figure 2: Precision@5 and Recall@5.

5 Results Discussion

Rule-based approaches, especially keyword-based methods, show the greatest limitations, even though they are fast and simple, their effectiveness depends on the exact word overlap between the question and the text content. The TF–IDF approach is an improvement to this, as it participates in capturing a broader lexical similarity, but is still sensitive to domain-specific words and does not understand semantic relationships between concepts.

Embeddings and the MiniLM-L6-v2 vector model, helps with semantic analysis. The precision is lower compared to TF–IDF which is expected, embeddings successfully capture meaning and thematic relatedness, but sometimes interpret semantic proximity too broadly, selecting parts that do not directly answer the question. This model achieves **recall** $=$ **1.0**, which means that it practically always identifies at least one relevant chunk for each question. However, it is important to emphasize that such a high recall *is not solely a result of the quality of the model*, but also a consequence of the way the evaluation is implemented. Specifically, the evaluation script uses the following binary classification criterion:

$$y_{pred} = (y_{scores} > 0),$$

which means that any positive score, regardless of its relative value or distribution, is interpreted as an indicator of “relevance”.

The supervised model shows a very similar pattern: it also achieves **recall = 1.0** due to the same evaluation mechanism. Logistic regression almost always assigns positive probabilities, so here too most chunks are classified as relevant according to the criterion ($p > 0$). The difference compared to the embedding model is seen in **Precision@5** and **Recall@5** where the supervised model achieves the best results, which makes it the most useful in the context of RAG systems. It is these metrics that reflect the real usability of the model: how well it ranks the most relevant segments at the very top of the list.

The difference compared to the embedding model is seen in **Precision@5** and **Recall@5** where the supervised model achieves the best results, which makes it the most useful in the context of RAG systems. Therefore, we treat the supervised classifier as our primary retrieval baseline for RAG, as it produces the most reliable top-k rankings.

The results confirm the hierarchy expected in information retrieval: rule-based models are the simplest and least capable of semantic understanding, embedding models introduce meaning but suffer from an overly broad interpretation of similarity, and the supervised model achieves the most stable performance thanks to learning from real examples. At the same time, the analysis shows that evaluation methods must be carefully designed, as inadequate thresholds or formulas can lead to misinterpretation of model quality.

In addition to quantitative metrics, a qualitative analysis was conducted, which can be found in `docs/` directory.

6 Conclusion

In Milestone 2, four baseline methods were successfully implemented and compared. The results show that each method contributes to understanding different aspects of the task of searching and ranking relevant content. Rule-based methods serve as minimal reference points, the embedding approach brings a powerful semantic component, while the supervised model achieves the most reliable ranking of key information that is of utmost importance in the RAG context. This work lays a foundation for Milestone 3, where the focus will shift to more advanced retrieval approaches and exploring improved strategies.