1 Implementation Process

1.1 Retriever Methods

- **B**M25 Retriever uses the term frequency (TF) and inverse document frequency (IDF) to rank documents.

$$\text{BM25}(D, Q) = \sum_{t \in Q} \text{IDF}(t) \cdot \frac{f(t, D) \cdot (k_1 + 1)}{f(t, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\overline{|D|}})}$$

- **S**emantic Retriever encodes documents and queries into dense vector embeddings and usually employs the cosine similarity metric which scores range from -1 to 1, defined as:

$$\text{Sim}(Q, D) = \frac{\mathbf{v}_Q \cdot \mathbf{v}_D}{\|\mathbf{v}_Q\| \|\mathbf{v}_D\|}$$

- **T**he Hybrid Retriever combines BM25's precision in keyword matching alongside the semantic depth of transformer-based embeddings to improve overall retrieval performance.

$$\text{Score}(D, Q) = \lambda \cdot \text{BM25}(D, Q) + (1 - \lambda) \cdot \text{Sim}(Q, D)$$

1.2 Evaluation Metrics

Use techniques in pre-retrieval, post-retrieval, and indexing stages to enhance the retrieval system based on BM25 and transformer with the ReAct reasoning chain.

- **Recall**: Recall calculates the ratio of gold-standard relevant evidence retrieved to the total gold evidence. It estimates how effectively the retrieval system is picking up relevant information. Higher recall indicates more relevant evidence is being retrieved.

$$Recall = \frac{|\text{Retrieved Evidence} \cap \text{Gold Evidence}|}{|\text{Gold Evidence}|}$$

- **Mean Reciprocal Rank (MRR)**: MRR measures the speed at which the first relevant result is obtained in the list. It provides the average of the reciprocal rank of the first suitable answer for every query. The correct answers are ranked higher earlier by the system.

$$MRR = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{rank_i}$$

- **Exact Match (EM)**: EM verifies whether the obtained answer exactly matches the gold-standard answer. It is a strict metric because slight differences (e.g., paraphrasing, synonyms) will result in a score of zero. High EM means high precision of answers.

$$EM = \frac{\sum_{i=1}^{N} \mathbb{1}(\text{Predicted Answer}_i = \text{Gold Answer}_i)}{N}$$

1.3 Additional Techniques

- **Pre-retrieval - Query Expansion**: Make the query expand to "Find documents about: query". By adding contextual information, the query intent can be more clearly defined, making the query more like a complete natural search language
- **Post-retrieval - Rerank and Fusion**: Retrieving documents separately from BM25 and semantic retrieval. Normalizing BM25 scores and then applying weighted fusion (bm25_weight * BM25_score + semantic_weight * Semantic_score) to rank documents by the combined score.
- **Indexing - Chunking Embedding**: Split large documents into smaller chunks for better retrieval with chunk_text(). The Semantic Retriever will convert text chunks into vector embeddings for semantic search.

2 Experiment Details & Result Analysis

2.1 Experiment 1

Experiment Setting & Result: Try to calculate evaluation metrics for BM25, Semantic, and Hybrid Retrieval. This table is the experiment result:

Type	Recall	MRR	EM
BM25 Only	61.67%	73.33%	6.67%
Semantic Only	46.67%	64.67%	13.33%
Hybrid (0.5 BM25 + 0.5 Semantic)	65.56%	86.67%	6.67%

Table 1: Evaluation Results for BM25, Semantic, and Hybrid Retrieval Models

Analysis:
BM25 gets high recall and MRR due to its keyword matching power. Despite Semantic Retriever havs higher EM, recall and MRR are lower and likely due to the fact that the method struggles with queries with small lexical variations. The equal-weight fusion (normalize the score of BM25) achieves higher recall and MRR than both individual methods while having the same EM as BM25.

This indicates that the combination of lexical and semantic representations results in more balanced retrieval performance.

2.2 Experiment 2

Experiment Setting & Result: Varying normalized BM25's proportions [0.1, 0.3, 0.5, 0.7, 0.9] in the Hybrid Model.

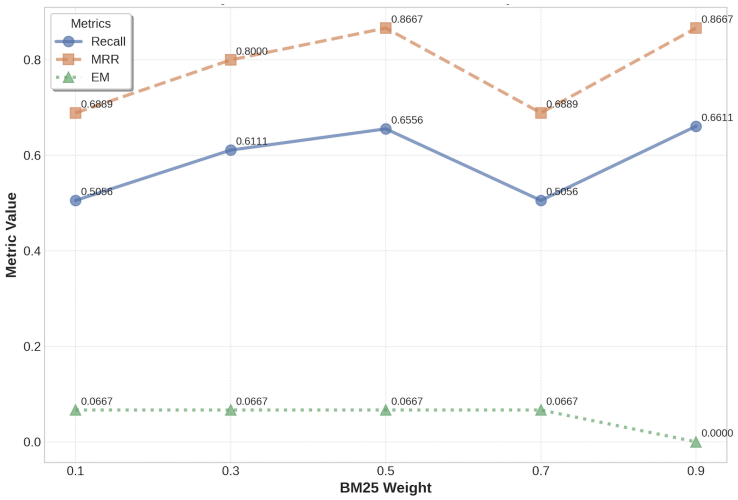

Figure 1: Hybrid Model Retrieval Metrics vs BM25 Proportion

Analysis:
Recall and MRR are both highest when BM25 weight is around 0.5–0.9, showing that the best performing hybrid overall is balanced or slightly biased towards BM25.

Exact Match (EM) remains constant at 0.0667 for most proportions, with a drop to 0.0 at 0.9 BM25. This can imply that giving too much weight to BM25 can sometimes come at the cost of losing the ability to recall exact matches, possibly by missing some semantic nuances. Or the 15 queries are not enough to show the correct trend.

3 Discussion & thoughts

Complementarity: The experiments show that the combination of BM25 and semantic retrieval techniques can outperform the use of either method by itself.

Limitations & Future Work: Future improvements can be incorporated in query routing, reranking methods and summarization to continue optimizing the retrieval process. Adaptive weighting methods based on query type or context can further enhance retrieval performance.