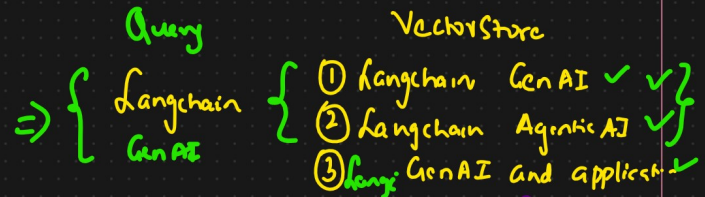


HYBRID SEARCH STRATEGIES

3) MMR [Maximal Marginal Relevance]

What is MMR?



MMR (Maximal Marginal Relevance) is a powerful diversity-aware retrieval technique used in information retrieval and RAG pipelines to balance relevance and novelty when selecting documents.

MMR selects documents that are both:

- 1. Relevant to the query ✓
 - 2. Diverse from each other (non-redundant) ✓
- } Aim

It prevents the retriever from returning very similar documents that repeat the same content.

$$MMR(d) = \lambda * \text{sim}(d, q) - (1 - \lambda) * \max_{s \in S} \text{sim}(d, s)$$

1) $Q_v = \text{Query}$

Here $\lambda \in [0, 1]$

2) Candidate document set D

Tunable parameter

3) Select Document(s)

1) Relevance to the q ($\uparrow \lambda$)

4) Similarity function $\text{sim}(a, b)$ (eg: cosine)

2) Diversity (low λ)

Problem

Query: How to use Langchain for RAG

Documents

D1

Langchain enable retrieval with FAISS

D2

Langchain can use Chroma and Pinecone too

D3

Langchain agents can call external tools

MMR

1) Pick Document \rightarrow highest similarity (doc, query)

<u>Doc</u>	<u>Similarity Score</u> (cosine similarity) $\text{Sim}(\text{doc}, Q)$
D1	0.95 ✓
D2	0.93
D3	0.80

1) Relevant Document \rightarrow D1.

Step 2: Select Second Document using MMR.

Compare D2 and D3

1) Relevance to query

2) ^{non}Redundancy with select Doc (D1)

$$\text{MMR}(D2) = \lambda * \text{Sim}(D2, q) - (1-\lambda) * \max_{S \in S} \text{Sim}(\text{doc}, S)$$

$\text{Sim}(\text{doc}, S)$

$\text{Sim}(D1, D2)$

$\text{Sim}(D1, D3)$

$\text{Sim}(D2, D3)$

Similarity Score

0.90 [Redundant]

0.30 [Diverse]

0.40

D1

$\lambda = 0.7$

$\text{Sim}(D1, D2)$

$$\text{MMR}(D2) = 0.7 * 0.93 + 0.3 * 0.90 = 0.651 - 0.27 = \underline{\underline{0.381}}$$

$$MMR(D_3) = 0.7 * 0.80 + 0.3 * Sim(D, D_3)$$

$$= 0.7 * 0.80 + 0.3 * 0.30 = 0.56 - 0.09 = \boxed{0.47} \uparrow\uparrow$$

$D_1 \rightarrow$ Relevance of the query ($Sim(d, q)$)

$D_3 \rightarrow$ MMR Score

Query

Rank	Document	Reason
1	D_1	Highest Relevant
2	D_3	But diversity + relevance [MMR] ←



When to Use MMR [Interviews]

1) In A RAG — To avoid feeding the LLM redundant documents

Eg: Langchain

1) FAISS 2) Agents 3) Memory 4) Prompt chain 5) Hybrid Retrieval

→ Result is richer, more useful input or context to LLM.

2) Chatbots : FAQ, SEARCH APP, DOCUMENT BROWSER

3) Retriever Already Return Many Result + Diversity

4) MMR + Hybrid Retrieval → Dense + Sparse



When Not to Use MMR

✖ Scenario

Extremely short context window

You need **precision only** ✓

Documents are already diverse

{ You're already reranking with LLM }

⊘ Why You Might Skip MMR

You may just want top-1 most relevant

Not focused on coverage

No need to enforce diversity

Redundancy handled by post-filtering