**Maximal Marginal Relevance (MMR)**

**🔹 What is MMR?**

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

**Aim:**

* Ensure documents are **relevant to the query**
* Ensure documents are **diverse from each other (non-redundant)**

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

**🔹 Formula**

MMR(d)=λ⋅sim(d,q)−(1−λ)⋅max⁡s∈Ssim(d,s)MMR(d) = \lambda \cdot sim(d, q) - (1-\lambda) \cdot \max\_{s \in S} sim(d, s)

* qq: Query
* DD: Candidate document set
* SS: Already selected documents
* sim(a,b)sim(a, b): Similarity function (e.g., cosine)
* λ∈[0,1]\lambda \in [0,1]: Tunable parameter
  + High λ → prioritize **relevance**
  + Low λ → prioritize **diversity**

**🔹 Example Walkthrough**

**Query:** *How to use LangChain for RAG*

Candidate Documents:

* D1: LangChain console retrieval with FAISS
* D2: LangChain can use Chroma/Pinecone
* D3: LangChain agents can call external tools

**Step 1:** Pick most relevant doc (highest similarity with query) → **D1**

**Step 2:** Select next doc using MMR → compare D2 & D3

* Relevance to query
* Redundancy with already selected doc (D1)

**Calculation Example:**

* λ = 0.7
* MMR(D2) = 0.7 × 0.93 + 0.3 × (redundancy) → 0.321
* MMR(D3) = 0.7 × 0.80 + 0.3 × 0.30 → 0.47 ✅

**Ranking Result:**

1. D1 (highest relevant)
2. D3 (adds diversity + relevance)

**🔹 When to Use MMR**

* In **RAG pipelines**: to avoid feeding redundant documents to LLMs
* With **LangChain** components:
  + FAISS, Agents, Memory, Prompt Chains, Hybrid Retrieval
* Scenarios:
  + Chatbots (FAQ, document search, knowledge assistants)
  + Search applications (enterprise search, e-discovery)
  + Document browsers (contracts, research papers, reports)

**Key Value:** Produces **richer, more diverse, and more useful context** for LLMs.

**🔹 When *Not* to Use MMR**

| **❌ Scenario** | **⭕ Why You Might Skip** |
| --- | --- |
| Short context window | You may just want top-1 most relevant |
| Precision only needed | Not focused on coverage |
| Documents already diverse | No need to enforce diversity |
| Already reranking with LLM | Redundancy handled by post-filtering |

**🔹 Real-Time Applications of MMR**

1. **Search Engines (Google, Bing, DuckDuckGo)**
   * Avoids showing users multiple search results with nearly identical content.
2. **E-commerce Product Search**
   * Ensures product recommendations are **relevant but varied** (e.g., different brands/models instead of duplicates).
3. **Recommendation Systems (Netflix, Spotify, Amazon)**
   * Prevents showing multiple items of the same type, balancing **user preference + diversity**.
4. **Legal & Compliance Document Retrieval**
   * Ensures retrieval surfaces both **highly relevant cases** and **novel precedents**.
5. **Healthcare & Research**
   * Retrieving patient records, studies, or trials without redundancy, but still covering diverse relevant cases.
6. **Conversational AI Assistants**
   * RAG pipelines where MMR ensures the LLM gets **diverse context documents** instead of repetitive info.

✨ **Final Takeaway:**  
MMR is a critical retrieval strategy in **hybrid search & RAG systems**, striking the right balance between **query relevance** and **document diversity**.