**Scenario: Reranking and Reciprocal Rank Fusion (RRF)**

You want to retrieve the most relevant documents for the query:  
**"What is machine learning?"**  
using query expansion (multi-query), RRF, and reranking.

**Step 1: User Sends Query with Query Expansion**

**Request:**

json

{  
 "query": "What is machine learning?",  
 "query\_expansion": {  
 "enabled": true,  
 "strategy": "multi\_query",  
 "llm\_config\_name": "nvidia-llama3-8b",  
 "num\_queries": 3  
 },  
 "use\_reranking": true,  
 "filter\_after\_reranking": true  
}

**Step 2: System Expands the Query**

**Expanded Queries:**

1. "What is machine learning?"
2. "Explain the basics of machine learning."
3. "How does machine learning work?"

**Step 3: Each Query Retrieves Documents**

**(Sample top-3 results per query before fusion):**

| **Query** | **Top Results (doc\_id, title)** |
| --- | --- |
| What is machine learning? | D1: "Intro to ML", D2: "ML Overview", D3: "History of ML" |
| Explain the basics of machine learning. | D2: "ML Overview", D4: "ML Basics", D5: "Supervised ML" |
| How does machine learning work? | D1: "Intro to ML", D5: "Supervised ML", D6: "ML Algorithms" |

**Step 4: RRF Fuses the Results**

**RRF Algorithm:**  
Documents appearing in multiple lists get a higher score.

**Merged & Scored:**

| **doc\_id** | **Titles** | **Appears in** | **RRF Score (example)** |
| --- | --- | --- | --- |
| D1 | Intro to ML | 1, 3 | 0.033 + 0.033 = 0.066 |
| D2 | ML Overview | 1, 2 | 0.025 + 0.033 = 0.058 |
| D5 | Supervised ML | 2, 3 | 0.020 + 0.025 = 0.045 |
| D4 | ML Basics | 2 | 0.020 |
| D3 | History of ML | 1 | 0.020 |
| D6 | ML Algorithms | 3 | 0.020 |

**(Top 3 after RRF: D1, D2, D5)**

**Step 5: Reranking**

The RRF-merged documents are sent to the reranker endpoint (e.g., Cohere API, model server).

**Sample reranker input:**

* Query: "What is machine learning?"
* Candidates: [D1, D2, D5]

**Reranker output (with scores):**

| **doc\_id** | **Title** | **Rerank Score** |
| --- | --- | --- |
| D2 | ML Overview | 0.95 |
| D1 | Intro to ML | 0.93 |
| D5 | Supervised ML | 0.82 |

**If**

filter\_after\_reranking=true

**and threshold is 0.85:**

* Only D2 and D1 are returned.

**Step 6: Final Response**

**Sample API Response:**

json

{  
 "results": [  
 {  
 "doc\_id": "D2",  
 "title": "ML Overview",  
 "score": 0.95,  
 "source\_queries": [  
 "What is machine learning?",  
 "Explain the basics of machine learning."  
 ]  
 },  
 {  
 "doc\_id": "D1",  
 "title": "Intro to ML",  
 "score": 0.93,  
 "source\_queries": [  
 "What is machine learning?",  
 "How does machine learning work?"  
 ]  
 }  
 ],  
 "query\_expansion\_metadata": {  
 "original\_query": "What is machine learning?",  
 "strategy": "multi\_query",  
 "requested\_num\_queries": 3,  
 "actual\_num\_queries": 3,  
 "expanded\_queries": [  
 "What is machine learning?",  
 "Explain the basics of machine learning.",  
 "How does machine learning work?"  
 ],  
 "total\_unique\_results": 6,  
 "query\_results\_summary": [  
 {  
 "query": "What is machine learning?",  
 "results\_count": 3,  
 "top\_similarity\_score": 0.95  
 },  
 {  
 "query": "Explain the basics of machine learning.",  
 "results\_count": 3,  
 "top\_similarity\_score": 0.95  
 },  
 {  
 "query": "How does machine learning work?",  
 "results\_count": 3,  
 "top\_similarity\_score": 0.93  
 }  
 ]  
 }  
}

**Summary**

* **Query Expansion** generates multiple related queries.
* Each query retrieves documents independently.
* **RRF** merges and scores all results, boosting those found by multiple queries.
* **Reranking** reorders the merged list for maximum relevance.
* **Filtering** (optional) removes low-scoring results.

This pipeline ensures you get the most relevant, diverse, and high-quality answers from your knowledge base.