**1) What Is an Ensemble Retriever?**

An **ensemble retriever** combines two or more search methods:

* **FAISS (Semantic)**: Finds documents similar in meaning (e.g. using neural embeddings).
* **BM25 (Lexical)**: Finds documents based on keyword matching (classic “bag-of-words” approach).

By merging both, you benefit from **semantic** understanding **and** exact keywords, often improving overall relevance.

**2) How It Works Under the Hood**

1. **Each Retriever** independently retrieves its top results, assigning each document a **rank** (1 = top, 2 = second, etc.).
2. These ranks are then combined using **Weighted Reciprocal Rank Fusion (WRRF)**.
   * We take each document’s rank from each system, convert it into a “score” (with higher weight for the system we trust more), and sum up for a final combined score.

The outcome is a **single** list of documents ordered by how well they rank across the two methods.

**3) Simple Example**

Imagine the query is **“air compressor”**, and each retriever returns five results:

**FAISS (Semantic)**

1. Rotary air compressor for factories
2. Heavy-duty compressor pump unit
3. Factory-grade air pumping machine
4. Compressed air solution – silent series
5. Energy-efficient rotary compressor

**BM25 (Lexical)**

1. Air compressor pump industrial use
2. Pump for air compression in industry
3. Industrial compressor air pump
4. Air pump for heavy industrial tools
5. Industrial use compressor model 1200

Let’s assign:

* **Semantic Weight** = **0.4**
* **Lexical Weight** = **0.6**

We then transform their ranks into a “score” and sum them. Here’s a small snippet:

| **Product** | **FAISS Rank** | **BM25 Rank** | **Final Score = 0.4 × FR + 0.6 × BR** |
| --- | --- | --- | --- |
| *Air compressor pump industrial use* | 5 | 1 | 0.4×5 + 0.6×1 = 2.6 |
| *Industrial compressor air pump* | 3 | 3 | 0.4×3 + 0.6×3 = 3.0 |
| *Rotary air compressor for factories* | 1 | 5 | 0.4×1 + 0.6×5 = 3.4 |
| *Heavy-duty compressor pump unit* | 2 | 6 | 0.4×2 + 0.6×6 = 4.4 |

*(Lower combined score indicates a higher overall rank in this example, so “Air compressor pump industrial use” wins.)*

In practice, the ensemble can incorporate more advanced formulas (like **reciprocal** of the ranks) and automatically handle overlaps. But the idea remains: **reward documents that rank well in both systems** while letting you control how much weight each approach has.

**Final Takeaway**

The **ensemble retriever** ensures you don’t miss relevant documents simply because they excel at either semantic matching or keyword matching alone. By weighting each approach, you get a **balanced** and often more **accurate** final ranking.