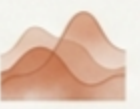

Hybrid Search: Fusing Keyword & Semantic Retrieval

*Bridging the gap between exact
matches and deep meaning.*



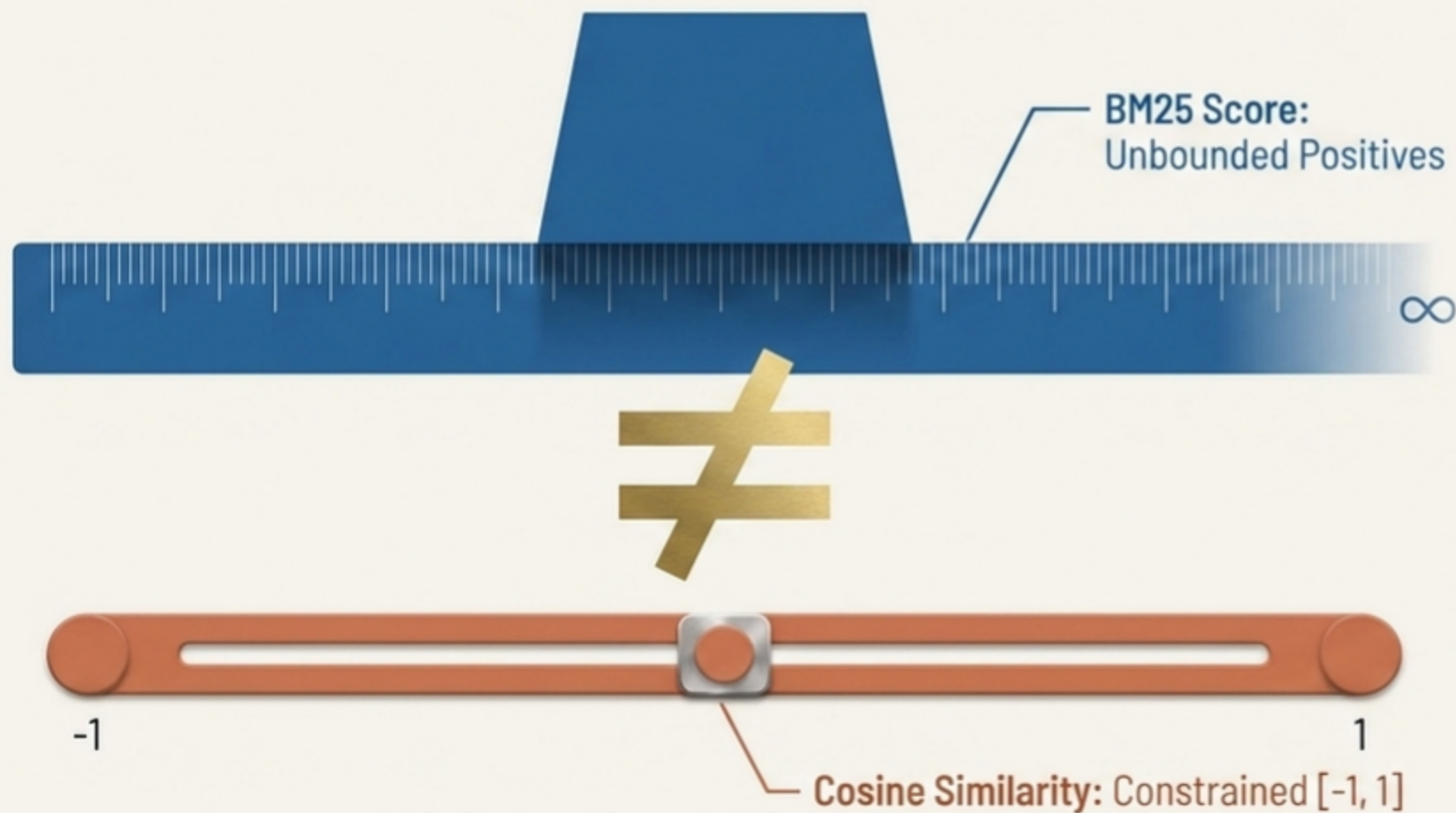
The Duality of Search: Complementary Failures

	 BM25 / Keyword Retrieval	 Dense / Semantic Retrieval
Core Strength	Nails exact words	Matches underlying meaning
Core Weakness	Misses conceptual context	Misses exact terms, names, and codes
Target	Product numbers, specific terminology	Fuzzy intent, related concepts

Where one fails, the other tends to succeed. The goal is synthesis.

The Scale Collision

Why raw scores refuse to mix.



Adding them lets BM25 dominate arbitrarily. It is mathematically equivalent to weighing an elephant and a feather on the same scale.



Reciprocal Rank Fusion (RRF)

The elegant, scale-free solution.

Sums across all retrievers
(BM25 + Dense).

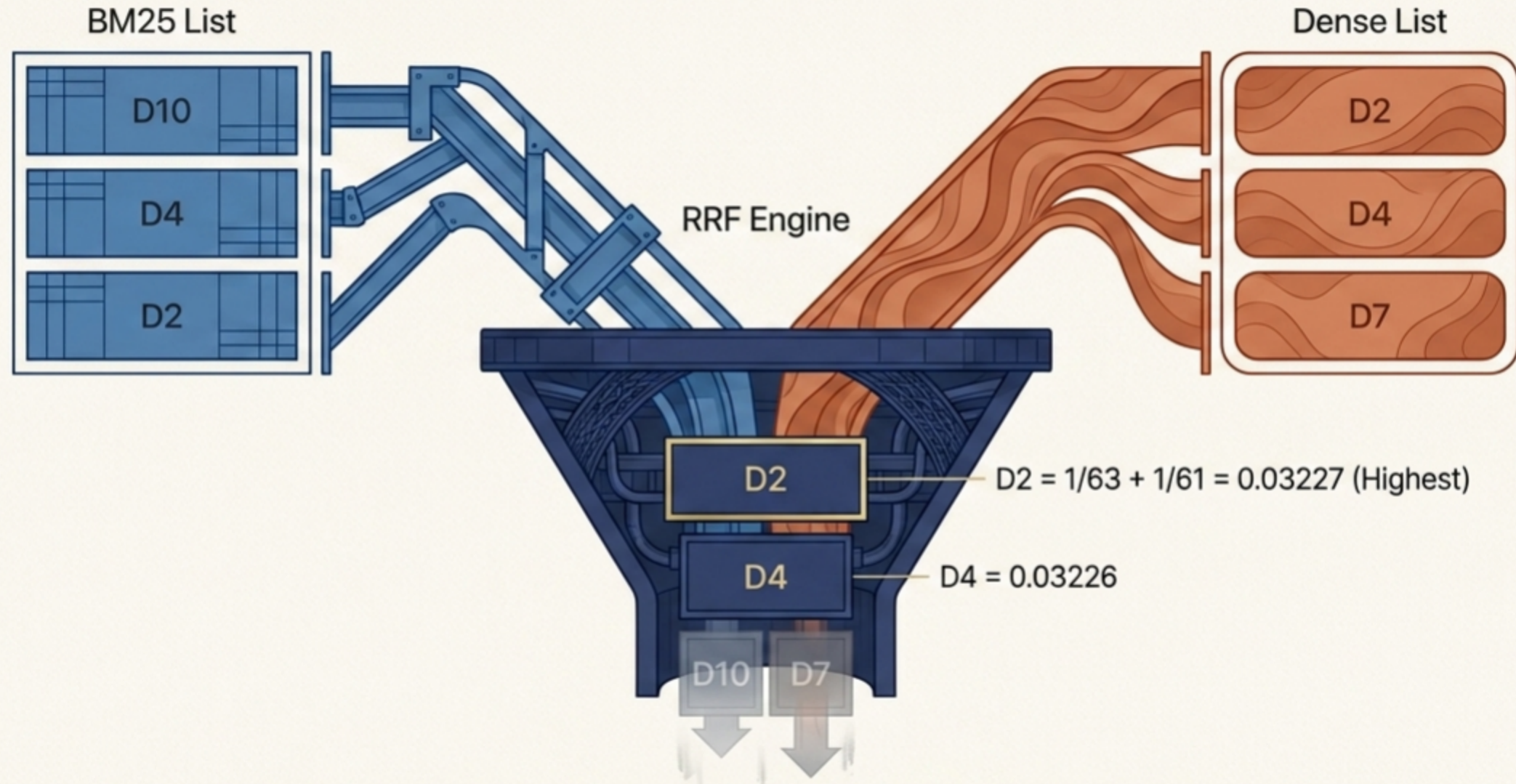
$$\text{RRF}(d) = \sum \left(\frac{1}{k + \text{rank}} \right)$$

A constant, typically ~60,
to smooth the distribution.

Uses the document's position
in the list, completely ignoring
the raw, incompatible scores.

Key Insight: By converting raw scores to relative rank, RRF becomes a robust, scale-free consensus mechanism in just a few lines of code.

RRF in Action: The Consensus Engine

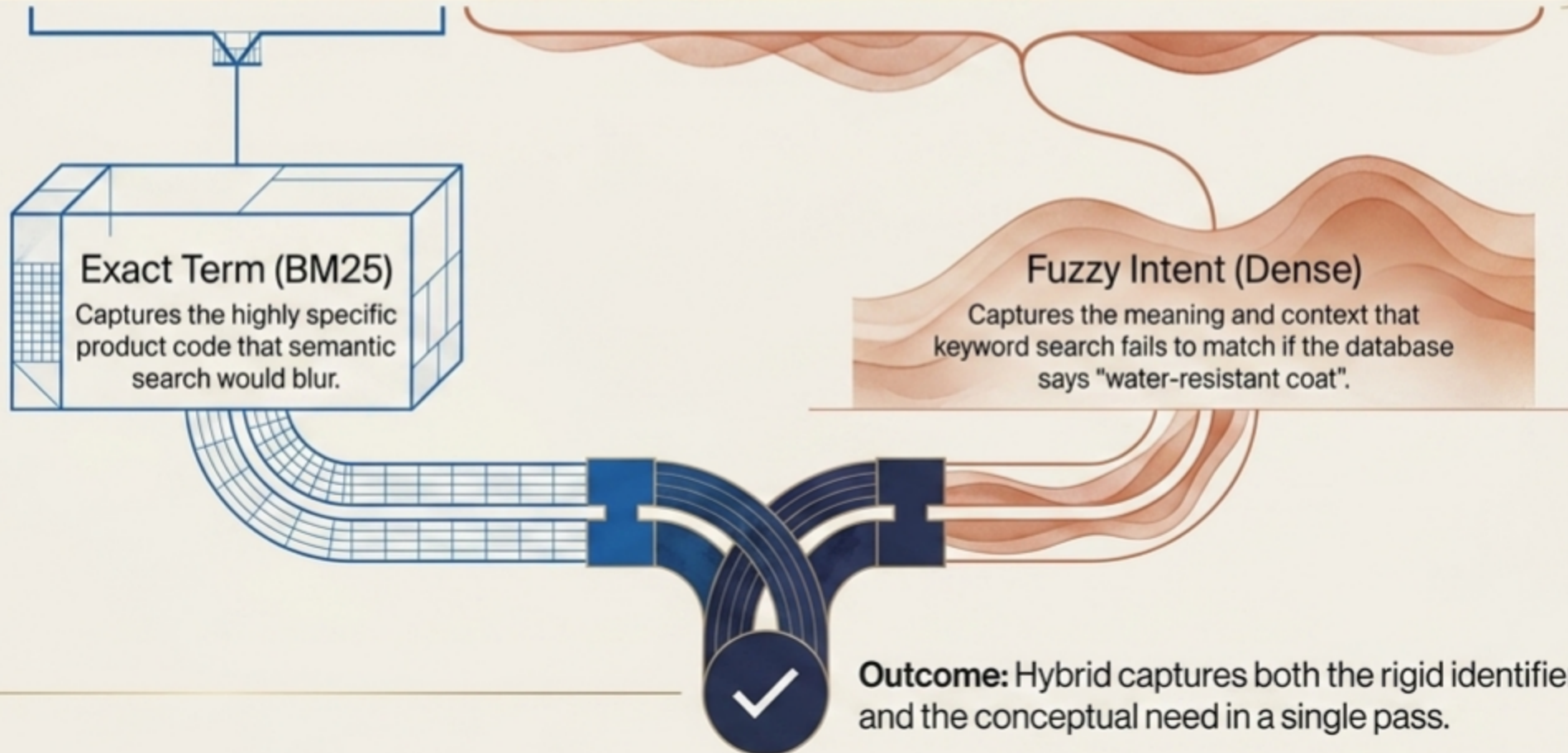


The Law of Consensus: Documents both retrievers like multiply in weight and rise to the top. Single-list-only orphans fall.

When Hybrid Wins

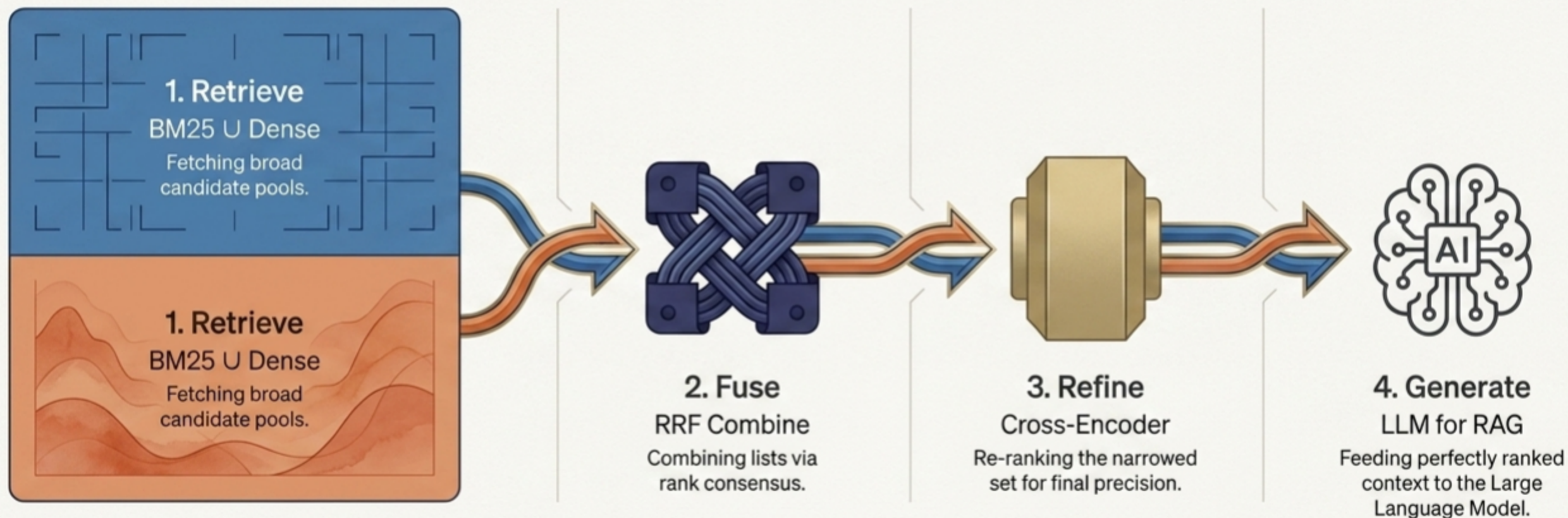
Capturing complex, multi-layered intent.

Q TX-900v2 durable waterproof jacket



The Full Modern Stack

Orchestrating retrieval for modern LLMs.



System Flow: Fusion creates the optimal candidate set;
Re-ranking sharpens it; the LLM articulates it.

Key Terms & Knowledge Check

Glossary

Hybrid Search

Combines multiple search techniques (e.g., keyword and semantic) to improve relevance by leveraging their complementary strengths.

Score Normalization

The process of scaling scores from different retrieval systems to a common range (e.g., 0 to 1) to allow for fair comparison and combination.

Reciprocal Rank Fusion (RRF)

A method for combining rankings by using the inverse of their rank position, emphasizing consensus across multiple search result lists without requiring score normalization.

Diagnostic

Q: Why can't you add a BM25 score and a cosine score?

A: They live on different scales; one dominates arbitrarily.

Q: What does RRF use instead of raw scores?

A: Each document's rank via $1/(k+\text{rank})$ — it relies on scale-free consensus.

Q: Where does re-ranking fit?

A: After fusion. Fuse to build the candidate set, then re-rank it.