Walmart Hybrid Retrieval – Deep Technical & Enterprise Guide

Final Option D Edition – with ASCII diagrams placed optimally for maximum clarity.

1. Executive Summary

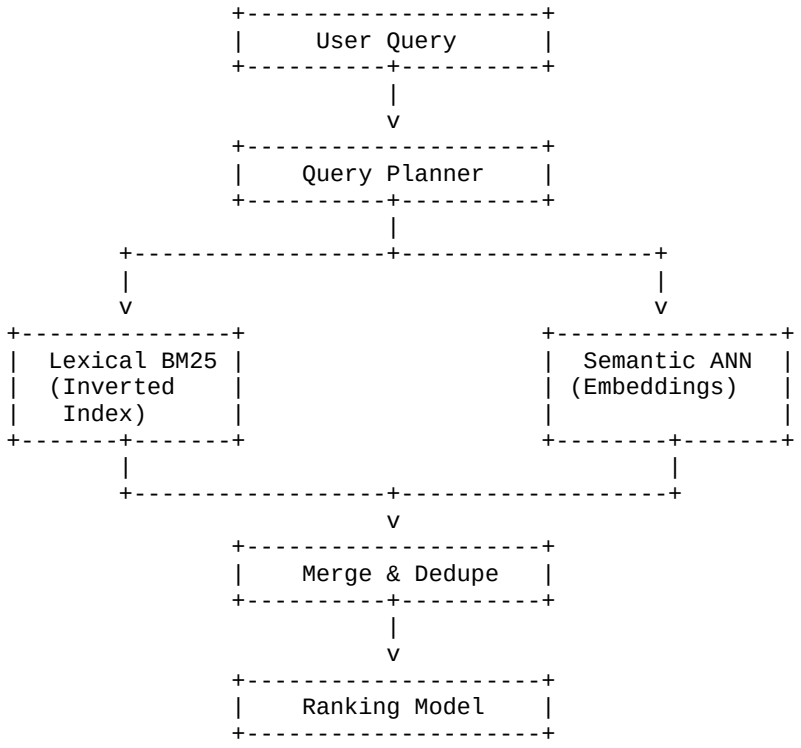
This document is your complete, merged Option■D guide: • Full deep technical explanation • Section■by■section conceptual unfolding • Enterprise mapping for Citi • AND all ASCII architecture diagrams inserted at optimal points This version is the final, polished edition.

2. What Walmart Needed to Solve

Walmart faces massive e■commerce search challenges: • Millions of SKUs • Sparse product titles • Tail queries with unusual phrasing • Vocabulary mismatch (sofa vs couch) • Extremely low latency requirements Lexical retrieval (BM25) is amazing for: • Exact terms • Numbers • Rare tokens • Identifiers Semantic retrieval (embeddings) is amazing for: • Understanding synonyms • Paraphrases • Intent Hybrid = both strengths.

Below is Walmart’s Hybrid Retrieval Architecture:

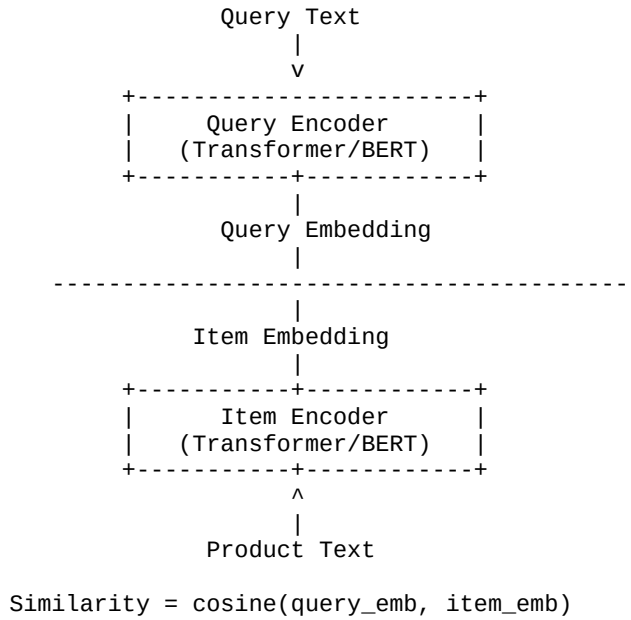
Diagram 1: Hybrid Retrieval Overview



3. Walmart’s Two■Tower Semantic Model

This is Walmart’s neural retrieval engine. The two■tower model encodes: • Query • Product …into the same embedding space, enabling cosine similarity search.

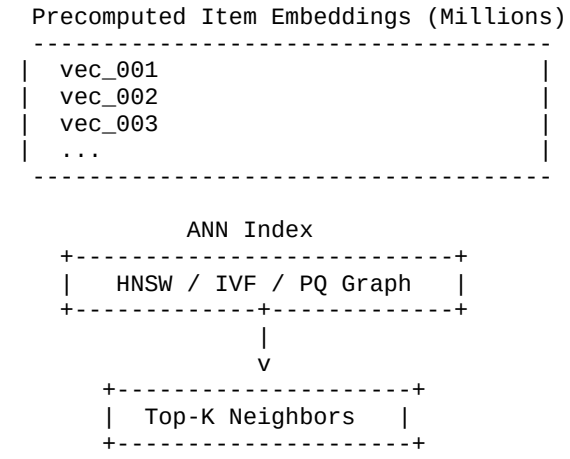
Diagram 2: Two■Tower Semantic Model



4. ANN Vector Search Layer

After product embeddings are computed, Walmart stores them inside an ANN index for fast lookup. This allows Walmart to search millions of embeddings in milliseconds.

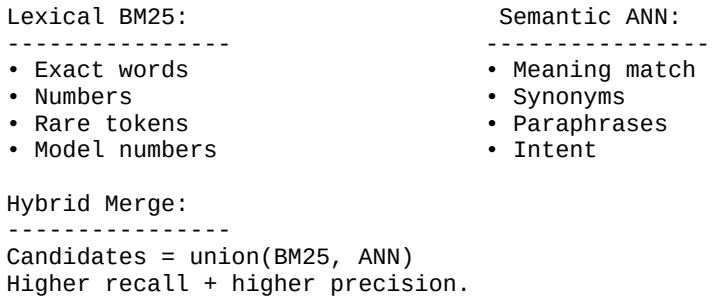
Diagram 3: ANN Vector Search



5. Hybrid Retrieval Logic – Why It Works

Hybrid retrieval is not two systems — it is one combined retrieval philosophy: Lexical handles exactness. Semantic handles meaning. Together: the highest recall + highest precision.

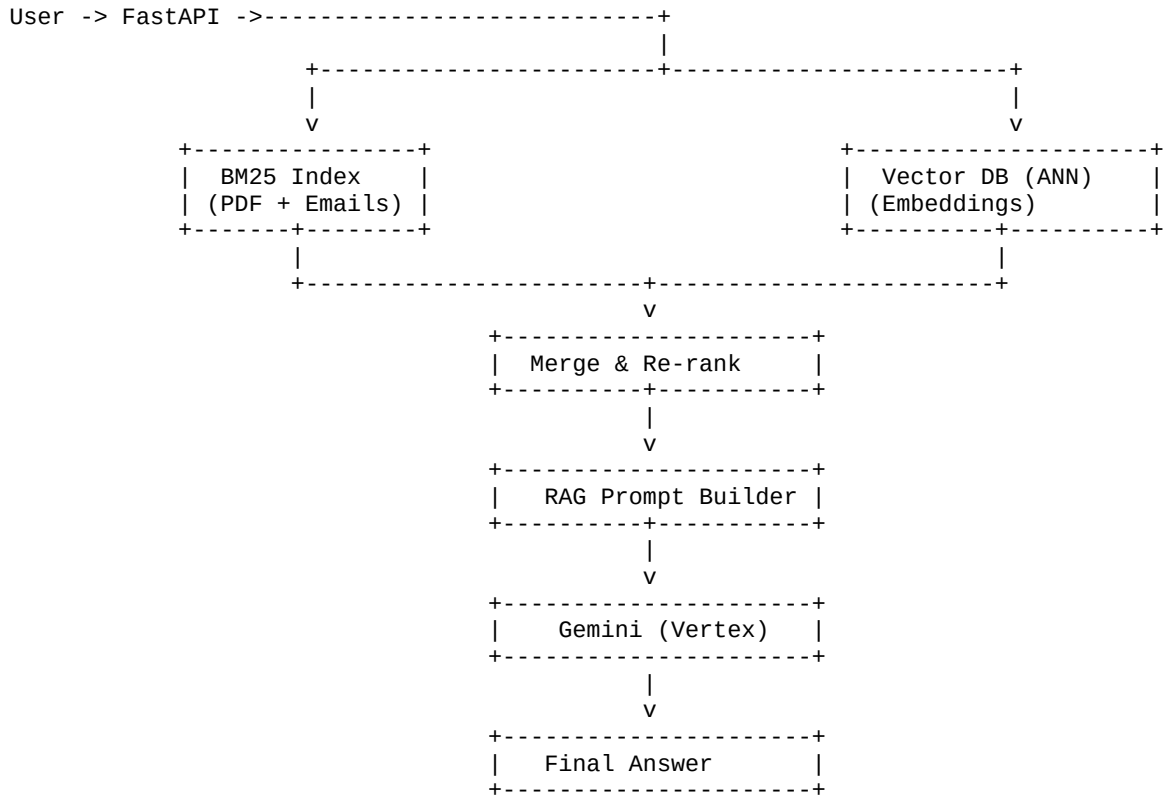
Diagram 4: Why Hybrid Wins



6. Mapping Walmart → Your Credit Copilot

Your architecture follows Walmart's blueprint precisely: • BM25 already built • RAG already built • Email + PDF ingestion already built • Vertex LLM already integrated All that's left is semantic embeddings + vector DB.

Diagram 5: Your Future Hybrid Credit Copilot Architecture



7. Final Notes

This is the fully merged, final Detailed PDF (Option■D) with all ASCII diagrams placed in their correct conceptual locations. You can use this PDF for: • Stakeholder demos • Architecture reviews • Design docs • Future planning for hybrid RAG