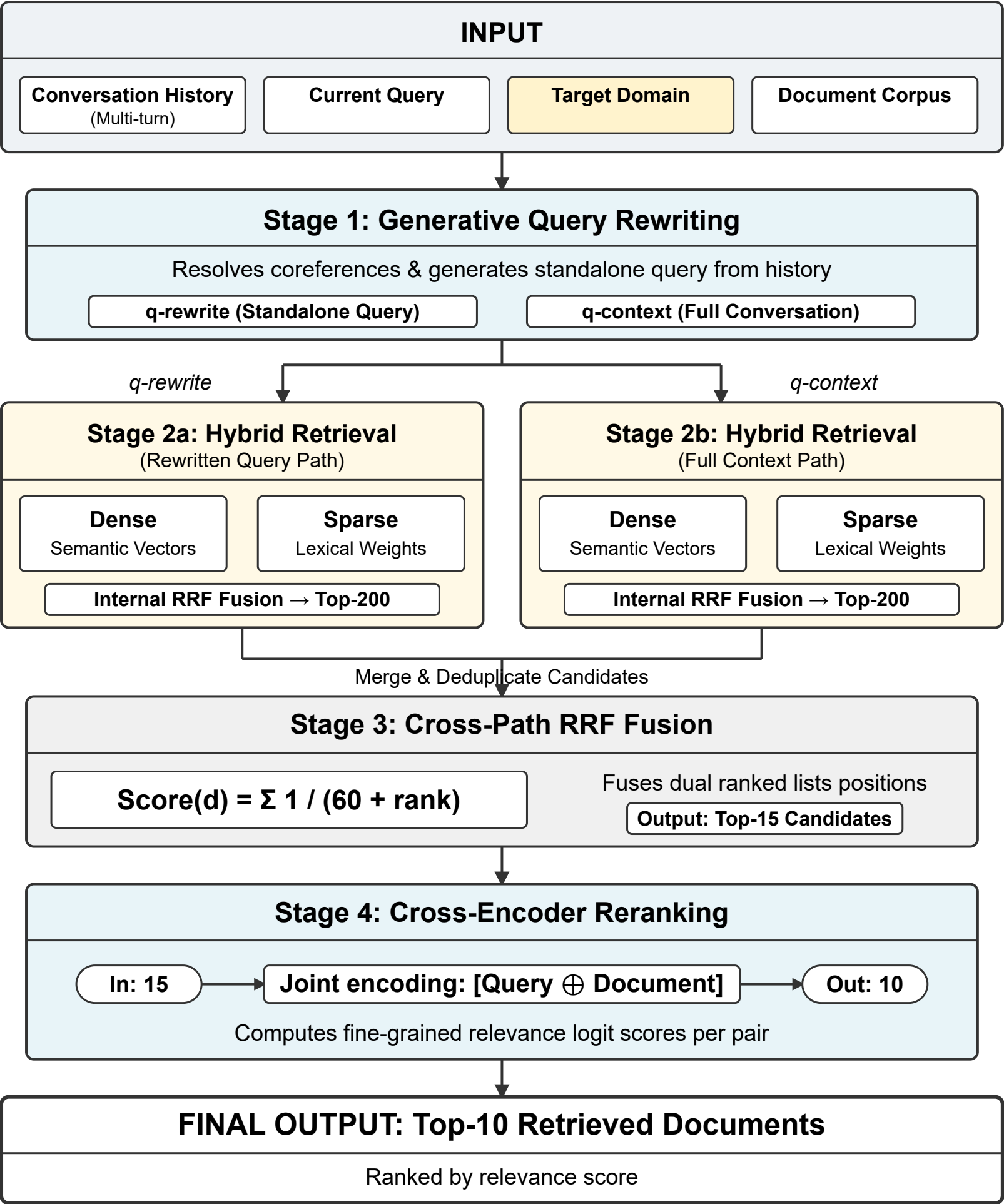


# Multi-Turn RAG System Architecture



## Models Used:

- Query Rewriter:** Qwen2.5-1.5B-Instruct (LLM, 1.5B params) — coreference resolution
- Embedding:** BAAI/bge-m3 (Bi-Encoder) — multi-functional dense+sparse vectors
- Reranker:** BAAI/bge-reranker-v2-m3 (Cross-Encoder) — joint attention scoring

