Multi-Turn RAG System Architecture

INPUT

Conversation History (Multi-turn) | *Current Query* (User Utterance) | *Target Domain* (e.g., Govt, FIQA) | *Document Corpus*

Stage 1: Generative Query Rewriting

Resolves coreferences & generates standalone query from conversation history

$q_{rewrite}$: Standalone Query | $q_{context}$: Full Conversation

$q_{rewrite}$ → Stage 2a

$q_{context}$ → Stage 2b

Stage 2a: Hybrid Retrieval

(Rewritten Query Path)

Dense Retrieval Semantic Vectors | **Sparse Retrieval** Lexical Weights

Internal RRF Fusion → Top-200

Stage 2b: Hybrid Retrieval

(Full Context Path)

Dense Retrieval Semantic Vectors | **Sparse Retrieval** Lexical Weights

Internal RRF Fusion → Top-200

merge & deduplicate candidates

Stage 3: Cross-Path RRF Fusion

RRF Scoring Formula

$Score(d) = \Sigma\ 1 / (60 + rank(r, d))$

Fuses dual ranked lists using rank positions

Output: Top-15 Candidates

Stage 4: Cross-Encoder Reranking

Input: 15 → *Joint encoding: [Query ⊕ Document]* → **Output: 10**

Computes fine-grained relevance logit scores per pair

Bi-Encoder: fast, approximate (Stage 2) | Cross-Encoder: slow, precise (Stage 4)

OUTPUT: Top-10 Retrieved Documents

Ranked by relevance score · Submitted as *submission.jsonl*

nDCG@5 = **0.1991** | +14.2% vs. initial run | Exceeds BM25 baseline

Models Used:

- **Query Rewriter:** Qwen2.5-1.5B-Instruct (LLM, 1.5B) coreference resolution and query standalone reformulation
- **Embedding:** BAAI/bge-m3 (Bi-Encoder) multi-lingual, multi-functional dense+sparse vectors (up to 8,192 t)
- **Reranker:** BAAI/bge-reranker-v2-m3 (Cross-Encoder) joint query-document attention scoring

- **Reranker:** BAAI/bge-reranker-v2-m3 (Cross-Encoder) joint query-document attention scoring