Phase 5: Enhanced RAG Implementation Report

Advanced Features Implemented

This phase enhanced the RAG pipeline with query rewriting and cross-encoder reranking. Query rewriting uses Flan-T5 to generate alternative question phrasings, expanding retrieval coverage. Cross-encoder reranking applies a pretrained MS-MARCO model to rescore retrieved passages, selecting the most relevant subset for answer generation.

Implementation Details

Query rewriting prompts Flan-T5 to generate alternative phrasings while preserving semantic meaning. The system retrieves passages for each variation and deduplicates results. Reranking uses cross-encoder/ms-marco-MiniLM-L-6-v2 to evaluate query-passage pairs jointly, then selects the top 5 from 10 initially retrieved passages.

Results

System Configuration	F1 Score	Exact Match
Baseline (top-10)	55.82%	50.0%
With Reranking Only	57.19%	52.0%
With Query Rewriting + Reranking	57.19%	52.0%

Analysis

Reranking provided a 1.37 F1 improvement (2.5% gain) by refining passage selection from 10 to 5 most relevant passages. Cross-encoder scoring successfully filtered noise from the initial retrieval pool, improving both F1 and exact match by 2 percentage points.

Query rewriting provided no additional benefit, achieving identical performance to reranking alone. Investigation revealed that Flan-T5 generated minimal variation—alternatives were often identical to originals. This implementation limitation prevented coverage expansion while adding computational overhead (21 minutes versus 17 minutes for reranking only).

The combined system's 1.37 F1 improvement stems entirely from reranking rather than query expansion. This demonstrates that implementation quality matters more than feature quantity—poorly executed advanced techniques add no value.

Key Findings

Reranking effectively improved precision by selecting higher-quality passages, justifying its 47% latency increase for batch processing scenarios. Query rewriting failed due to insufficient output diversity from the generative model. Production deployment should implement reranking while excluding query rewriting until reformulation quality improves. Future work should explore alternative query expansion methods such as keyword-based expansion or embedding interpolation rather than generative rewriting.