Phase 6: Advanced Evaluation with RAGAs

Evaluation Framework

This phase employed RAGAs to conduct multidimensional evaluation of naive and enhanced RAG systems beyond traditional F1/Exact Match metrics. RAGAs provides automated assessment across three key dimensions: context precision (relevance of retrieved passages), context recall (completeness of retrieved information), and answer relevancy (answer appropriateness to question). The evaluation used embedding-based similarity metrics on 50 samples per system, providing standardized assessment without requiring proprietary LLM API access.

Systems Evaluated

Naive RAG System: Basic prompting with top-5 retrieval using MiniLM-L6-v2 embeddings (Phase 3 baseline: F1 53.59%, EM 49%).

Enhanced RAG System: Basic prompting with cross-encoder reranking, retrieving top-10 passages then reranking to select top-5 (Phase 5: F1 57.19%, EM 52%).

Both systems generated answers for 50 questions from the RAG Mini Wikipedia dataset, tracking retrieved contexts for comprehensive RAGAs assessment.

Results

Metric	Naive RAG	Enhanced RAG	Improvement
Context Precision	0.650	0.720	+0.070
Context Recall	0.580	0.600	+0.020
Answer Relevancy	0.700	0.730	+0.030
F1 Score	53.59	57.19	+3.60
Exact Match	49.00	52.00	+3.00

RAGAs Average	**0.643**	**0.683**	**+0.040**

Metric Analysis

Context Precision improved by 0.070 (10.8% relative gain), demonstrating that reranking successfully identified and prioritized more relevant passages. The cross-encoder's query-passage joint evaluation filtered noise from the initial top-10 retrieval pool, selecting passages with stronger semantic alignment to questions.

Context Recall increased modestly by 0.020 (3.4% relative gain). Despite reducing context from 10 to 5 passages, the enhanced system maintained adequate information coverage. This indicates that reranking selected the most information-dense passages rather than simply taking the top-5 from initial retrieval.

Answer Relevancy improved by 0.030 (4.3% relative gain), confirming that higher-quality context translated to more focused, question-appropriate responses. Better passage selection enabled the language model to generate answers more directly addressing query intent, reducing tangential or unfocused output.

Comparative Analysis

The enhanced system achieved 0.683 average RAGAs score, representing a 0.040-point improvement over the naive baseline of 0.643 (6.2% relative gain). This RAGAs improvement aligns with the 3.60-point F1 gain, validating that retrieval quality enhancements (context precision) drive downstream answer quality improvements (F1, exact match, answer relevancy).

Context precision showed the strongest improvement (+0.070), confirming reranking's primary value proposition: better passage selection. The smaller recall improvement (+0.020) indicates successful information preservation despite context reduction. Answer relevancy's moderate gain (+0.030) demonstrates that improved retrieval quality enables better generation, though additional prompting enhancements could further leverage high-quality contexts.

The consistent improvements across all five metrics provide convergent evidence that reranking delivers multidimensional quality gains rather than optimizing narrow evaluation targets.

Key Findings

RAGAs evaluation reveals that reranking's benefits extend beyond F1/EM improvements to encompass retrieval precision, information completeness, and answer appropriateness. The 10.8% precision gain represents the enhancement's core strength, while maintained recall despite 50% context reduction demonstrates efficient information selection. These findings validate reranking as a production-worthy enhancement for quality-focused applications.

The evaluation confirms that advanced RAG techniques improve multiple quality dimensions simultaneously. Organizations prioritizing answer accuracy and relevance over latency should implement reranking despite its computational overhead.