Ground the Domain: From Naïve RAG to Production Patterns

Phases 3, 4 and 6 Evaluation Report

Chongdi Wang

I Step 3: Naïve RAG Evaluation

The goal of this evaluation phase was to benchmark a naïve retrieval-augmented generation (RAG) pipeline using different prompting strategies. The setup followed the assignment guidelines: each query retrieved a single top-1 evidence passage from the Milvus vector index, and the answer was generated by the Flan-T5 model hosted in Google Colab. Performance was measured using the Hugging Face SQuAD metrics, reporting both Exact Match (EM) **and** F1 scores.

Three prompting strategies were tested:

(1) *Instructional prompt*. "Answer STRICTLY using the context. If insufficient, reply 'I don't know.'"

(2) *Chain-of-Thought (CoT) prompt*. "You are a careful analyst. Use ONLY the context. If insufficient, say 'I don't know.' Carefully plan your reasoning step by step."

(3) *Persona prompt.* "You are a concise encyclopedia editor. Use ONLY the context. If insufficient, say 'I don't know.' Keep answers factual and brief."

Initial sanity checks on a subset (20 samples) showed widely varying performance: CoT prompts returned near-zero EM, while Instruction and Persona prompts achieved moderate scores (around 35% EM and ~44 F1). This indicated that while CoT elicited more verbose reasoning, it introduced substantial mismatch with the short ground-truth answers used in evaluation.

Subsequent evaluation on the full dataset confirmed these patterns. The Persona strategy consistently outperformed the two alternatives, yielding the highest F1 (31.2) and EM (23.9). The Instruction prompt ranked second with F1 ≈ 28.5 and EM ≈ 22.0, while the CoT prompt performed poorly with F1 ≈ 9.8 and EM ≈ 2.4. The results highlight the sensitivity of automatic metrics to output style: since SQuAD metrics rely on lexical overlap, verbose or explanatory responses (as seen in CoT) can reduce alignment with concise reference answers, even if factually correct. By contrast, the Persona framing naturally constrained the model to concise, fact-style responses that more closely matched the gold labels.

In summary, the evaluation demonstrates that prompt design significantly affects measured performance in RAG. For this dataset, concise context-grounded prompts are superior**,** while free-form reasoning hurts automatic metrics. This outcome establishes a clear baseline for Step 3 and motivates subsequent experiments (Step 4) to test parameter variations (e.g., top-k evidence size) and later enhancements (Step 5) such as reranking and citation grounding.

II Step 4: Paremeter Experimentation

The objective of Step 4 was to systematically assess how retrieval depth (*top-k*), reranking strategies, and embedding dimensionality affect retrieval-augmented generation (RAG) performance. The evaluation used two sentence-transformer models with different embedding sizes—all-MiniLM-L6-v2 (384 dimensions) and all-mpnet-base-v2 (512 dimensions)—combined with two selection strategies--*concat* and Maximal Marginal Releveance (*MMR)* and three retrieval depths (*k = 3, 5, 10*). The LLM used was Flan-T5 transformer with a persona-style prompt, and performance was measured with Exact Match (EM) and F1 using Huggingface SQuAD.

Key Observations.

1. **Embedding Model Effects.**
   Overall, the MiniLM6v2 model with concat consistently delivered the best performance, achieving its peak at *top-k=5* with an F1 of 53.3 and EM of 48.0. This outperformed mpnetv2, even though mpnetv2 has a larger embedding size (512 vs 384). The result suggests that, for this dataset and chunk size, a lighter embedding model can be competitive or superior, possibly due to more compact representations that align better with retrieval granularity.
2. **Impact of Retrieval Depth (*top-k*).**
   The results reveal a non-monotonic relationship between *top-k* and F1. For MiniLM6v2-concat, increasing *k* from 3 to 5 improved F1 (52.1 → 53.3), but performance dropped at *k=10* (50.1). Similarly, mpnetv2-concat peaked at *k=3* (52.8 F1) and degraded with higher *k*. This indicates that larger *k* values might increase noise and redundancy in context that dilutes answer quality. Retrieval should therefore balance recall with context precision, as excessive evidence can overwhelm the generator.
3. **Reranking Strategy (MMR vs Concat).**
   Across both models, the MMR strategy underperformed relative to simple concatenation. For MiniLM6v2, MMR scored substantially lower (41.2–38.8 F1 range) compared to concat (50–53 F1). Mpnetv2 displayed a similar pattern, with concat outperforming MMR by ~3–7 points. This suggests that in this setting, MMR's diversification disrupted the tight alignment between query and top passages, leading to weaker grounding. While MMR can be useful for diverse or redundant corpora, here it reduced retrieval relevance.
4. **Exact Match vs F1 Trends.**
   Both metrics moved in tandem, with MiniLM6v2-concat achieving the highest EM

of 48.0, aligning closely with its F1 peak. This reinforces that high-scoring configurations improved both lexical overlap and partial answer correctness, demonstrating consistency in quality.

Interpretation and Implications. The findings indicate that model choice and retrieval configuration strongly affect RAG effectiveness. For this dataset, a smaller embedding model with careful retrieval depth outperformed a larger model. The diminishing returns of larger *k*highlight the risk of "context dilution," while the MMR results caution against over-diversification in smaller corpora.

These experiments highlight the importance of tuning retrieval parameters to dataset characteristics. The optimal setting—MiniLM6v2 with concat and *top-k=5*—establishes a strong baseline. Future work should explore adaptive top-k strategies, reranker models (e.g., cross-encoders), and longer-context generators to mitigate noise from larger retrieval depths.

III Step 6: Advanced Evaluation Using RAGAS

The evaluation phase focused on systematically assessing the performance of two retrieval-augmented generation (RAG) pipelines: a *naive baseline* and an *advanced configuration*. To provide an interpretable and replicable benchmark, the study employed RAGAS (Retrieval Augmented Generation Assessment Suite), a toolkit designed for structured evaluation of retrieval-generation systems. RAGAS offers both LLM-based and non-LLM metrics, but due to compatibility and resource constraints, the current implementation emphasizes ID-based context precision and recall. These metrics directly measure whether retrieved passages contain the ground-truth evidence necessary to answer a question.

Evaluation Approach. The assessment followed three main steps. First, outputs from the naive and advanced pipelines were aligned with the *rag-mini-wikipedia* dataset of factoid question–answer pairs. Each pipeline returned candidate answers alongside the supporting contexts retrieved from a Milvus/FAISS index. Second, passages were mapped to document identifiers, enabling reliable computation of retrieval hit-rates. Third, evaluation metrics were computed using set-based overlap measures:

- *Context Precision*: the proportion of retrieved passages that are relevant.
- *Context Recall*: the proportion of relevant passages successfully retrieved.
  Both macro (average over questions) and micro (global counts) scores were reported. This design ensures robustness against individual outliers while maintaining sensitivity to aggregate trends.

Results. The evaluation used 50 sampled questions. The naive pipeline achieved macro and micro precision of 0.16, with a corresponding recall of 0.16. By contrast, the advanced pipeline showed markedly different behavior: precision dropped to 0.048, while recall increased to 0.24. In other words, the advanced configuration retrieved more of the ground-truth passages in general, but did so with considerably lower selectivity. This trade-off is

visible in both the macro and micro statistics, which track closely and confirm consistency across the sample.

Interpretation. The results highlight a fundamental tension between breadth of retrieval **and** quality of retrieval**.** The naive pipeline returns fewer passages overall but maintains higher precision, suggesting that its retrieved evidence is more tightly aligned with the gold reference. The advanced pipeline, likely due to more aggressive retrieval parameters and re-ranking, increases coverage but introduces noise, which lowers precision. In downstream applications, this difference has significant implications: low recall risks missing critical evidence, while low precision burdens the generator with irrelevant information, potentially degrading answer quality. Striking a balance between these metrics remains a central challenge in RAG optimization.

Conclusion. This phase demonstrates the value of RAGAS evaluation in isolating retrieval quality independently of generative fluency. The observed trade-off between naive and advanced pipelines underscores the need for parameter tuning and potentially hybrid strategies (e.g., multi-stage retrieval followed by context filtering) in later phases. Overall, the report establishes a baseline for comparing further iterations and validates the use of precision-recall analysis as a reliable proxy for answer validity.