Analysis of Pre- and Post-Rerank Performance

Common Context

- **Questioning LLM:** Qwen/Qwen3-1.7B
- **Judge LLM (Faithfulness & Relevance):** LLaMA3-70B-8192
- **Temperature:** 0.4
- **Retriever:** deepseek-r1-distill-llama-70b
- **Translation Length:** 473
- **Reference Length:** 682

What Stayed the Same

- **Faithfulness & Relevance (LLM-rated):** Both remained at 1.0, indicating perfect factual and topical alignment.
- **Lexical Overlap (ROUGE and BLEU):** No change in ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum, or BLEU score. This suggests surface-level similarity with the reference answer remained unchanged.

What Decreased

- **BERTScore (Semantic Similarity):**
  - Precision: decreased from 0.8190 to 0.7949
  - Recall: decreased significantly from 0.8270 to 0.6866
  - F1 Score: decreased from 0.8230 to 0.7368

  Interpretation: The post-rerank responses were more concise but lost some semantic alignment with the reference.
- **RAGAS Context Precision:** Dropped from 0.8056 to 0.7500
  Interpretation: The re-ranked chunks were slightly less precise in relevance, possibly introducing generalization or noise.

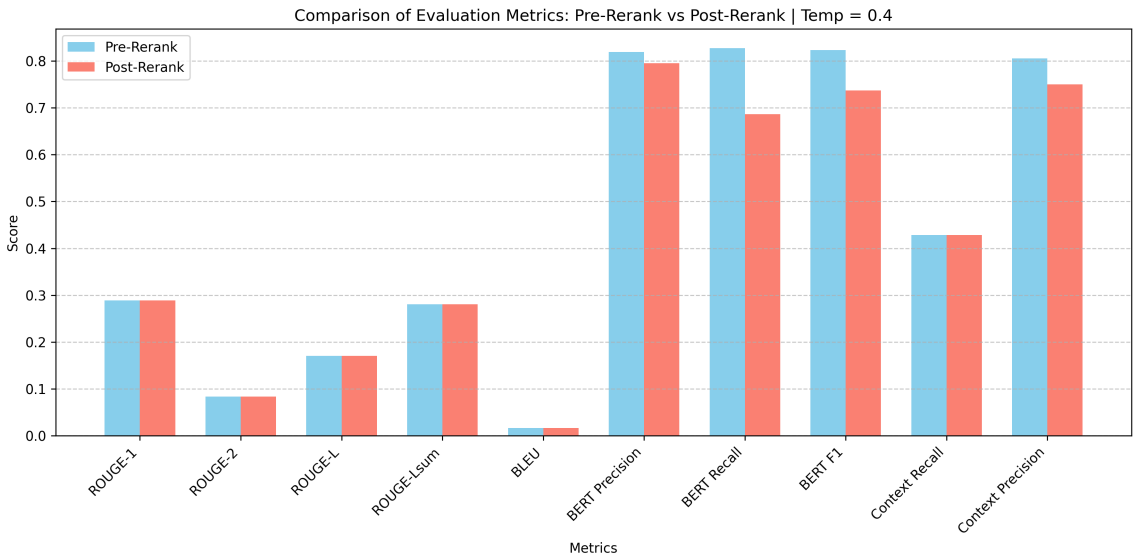

Figure 1: Enter Caption

What Didn't Improve

Despite reranking:

- No change in ROUGE or BLEU indicates that answer fluency or token-level overlap did not improve.
- Minor declines in semantic similarity and contextual precision suggest reranking didn't yield a net quality gain.

Key Takeaways

- Reranking maintained high factual consistency and topic relevance.
- However, it did not enhance lexical or semantic metrics and even slightly reduced semantic richness and precision.
- This suggests that reranking, in this setup, may prioritize structurally coherent but semantically narrower content.
- **Recommendation:** Consider hybrid reranking strategies using confidence scores, query-document alignment, or model-based scoring to better preserve both precision and semantic depth.