Advanced RAG Pipeline

So, for the advanced RAG, I implemented two advanced features in the RAG:
1. Query Rewriting
2. Reranking

Query Rewriting Implementation:
The enhancement uses the existing Flan-T5 model to generate alternative phrasings of the query, expanding the search space.

- Generates 2 alternative phrasings using Flan-T5
- Retrieves passages for each query variant
- Expands coverage beyond exact keyword matching

Reranking Implementation
Uses a cross-encoder to re-score retrieved passages based on semantic relevance to the original query (not the rewrites).

- Over-retrieves 10 passages per query variant
- Scores semantic relevance using ms-marco-MiniLM cross-encoder
- Selects the top 3 most relevant for generation

Evaluation Methodology

- Exact Match: Character-for-character match with ground truth
- Fuzzy Match: Allows substring/containment matches
- Average F1: Sequence similarity score (0-100%)

Results:

```
Results
| Pipeline                      |   Exact Match (%) |   Fuzzy Match (%) |   Avg F1 (%) |
|:------------------------------|------------------:|------------------:|-------------:|
| Naive RAG (Baseline)          |           44.6667 |           59.3333 |      56.5602 |
| Advanced RAG (Rewrite+Rerank) |           46.6667 |           62      |      59.4995 |

IMPROVEMENTS:
Exact Match: +2.0 percentage points
Fuzzy Match: +2.7 percentage points
Average F1: +2.9 percentage points

Advanced_rag_comparison.csv
Saved advanced_rag_detailed_comparison.csv
Summary

Results:

Naive RAG (Baseline):
   - Exact Match: 44.7%
   - Fuzzy Match: 59.3%
   - Average F1: 56.6%

Advanced RAG (Rewrite + Rerank):
   - Exact Match: 46.7%
   - Fuzzy Match: 62.0%
   - Average F1: 59.5%
```

RAGAs

RAGAs is an automated evaluation framework specifically designed for retrieval-augmented generation systems. Unlike traditional metrics (Exact Match, F1) that perform string-matching between predictions and ground truth, RAGAs employs a large language model (gpt-4o-mini in this implementation) as an impartial judge to assess semantic quality across four orthogonal dimensions:
Metrics:

- Faithfulness
- Answer Relevancy
- Context Precision
- Context Recall

Metrics Explained:
1. Faithfulness (0-1)
   Measures if the answer is grounded in the retrieved contexts.
   Contexts support higher-level = answer statements.

2. Answer Relevancy (0-1)
   Measures if the answer addresses the question.
   Higher = answer is semantically aligned with the question.

3. Context Precision (0-1)
   Measures if relevant contexts rank highly in retrieval.
   Higher = important contexts appear at the top.

4. Context Recall (0-1)
   Determines if contexts contain the necessary information for the answer.
   Higher = all necessary information was retrieved.

Results from RAGAs

```
Successful evaluations:
  Naive RAG: 150/150 queries
  Advanced RAG: 150/150 queries

Final RAGAs Comparison
|              |   Faithfulness |   Answer Relevancy |   Context Precision |   Context Recall |
|:-------------|---------------:|-------------------:|--------------------:|-----------------:|
| Naive RAG    |       0.704444 |           0.686802 |            0.671667 |              0.6 |
| Advanced RAG |       0.732222 |           0.719409 |            0.843333 |         0.633333 |
| Improvement  |       0.0277778 |          0.0326076 |            0.171667 |         0.0333333 |

DETAILED METRIC BREAKDOWN
```

```
RAGAS METRICS:

Naive RAG (Baseline):
  Faithfulness:      0.7044
  Answer Relevancy:  0.6868
  Context Precision: 0.6717
  Context Recall:    0.6000
  Average Score:     0.6657

Advanced RAG (Query Rewriting + Reranking):
  Faithfulness:      0.7322
  Answer Relevancy:  0.7194
  Context Precision: 0.8433
  Context Recall:    0.6333
  Average Score:     0.7321

Performance Change:
  Overall: +0.0663 (+6.6%)
```