

RAG Evaluation Results

Evaluation Date: December 15, 2025

Evaluation Time: 12:00:38

Overview

This document summarizes the evaluation of predicted answers against actual answers using three key metrics:

- **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation)
- **BLEU** (Bilingual Evaluation Understudy)
- **Semantic Similarity** (Cosine Similarity)

Evaluation Results by Pair

Pair 1: Capital of France

Actual: The capital of France is Paris.

Predicted: Paris is the capital city of France.

Metric	Precision	Recall	F1 Score
ROUGE-1	0.8571	1.0000	0.9231
ROUGE-2	0.3333	0.4000	0.3636
ROUGE-L	0.5714	0.6667	0.6154

- **BLEU Score:** 0.0831
- **Semantic Similarity:** 0.9698 ⭐

Pair 2: Great Wall of China

Actual: The Great Wall of China is in Beijing.

Predicted: The Great Wall is located in China, near Beijing.

Metric	Precision	Recall	F1 Score
ROUGE-1	0.7778	0.8750	0.8235
ROUGE-2	0.2500	0.2857	0.2667
ROUGE-L	0.6667	0.7500	0.7059

- **BLEU Score:** 0.1411
- **Semantic Similarity:** 0.9269

Pair 3: Water Boiling Point

Actual: Water boils at 100 degrees Celsius.

Predicted: At 100°C, water starts to boil.

Metric	Precision	Recall	F1 Score
ROUGE-1	0.5714	0.6667	0.6154
ROUGE-2	0.1667	0.2000	0.1818
ROUGE-L	0.2857	0.3333	0.3077

- **BLEU Score:** 0.0485
- **Semantic Similarity:** 0.9174

Pair 4: Python Programming Language

Actual: Python is a popular programming language.

Predicted: Many developers use Python as a programming language.

Metric	Precision	Recall	F1 Score
ROUGE-1	0.5000	0.6667	0.5714
ROUGE-2	0.1429	0.2000	0.1667
ROUGE-L	0.5000	0.6667	0.5714

- **BLEU Score:** 0.0699
- **Semantic Similarity:** 0.8572

Summary Statistics

Average Scores

Metric	Average Score
ROUGE-1 F1	0.7334
ROUGE-2 F1	0.2447
ROUGE-L F1	0.5501
BLEU	0.0857
Semantic Similarity	0.9178

Key Insights

1. **Highest Performing Pair:** Pair 1 (Capital of France)
 - Best ROUGE-1 F1: 0.9231
 - Best Semantic Similarity: 0.9698
2. **Lowest Performing Pair:** Pair 4 (Python Programming Language)
 - Lowest ROUGE-2 F1: 0.1667
 - Lowest Semantic Similarity: 0.8572
3. **Metric Analysis:**
 - **ROUGE-1** scores are consistently high (0.57-0.92), indicating good unigram overlap
 - **ROUGE-2** scores are moderate (0.17-0.36), showing less bigram matching
 - **BLEU** scores are low (0.05-0.14), which is expected for short sentences with structural differences
 - **Semantic Similarity** scores are excellent (0.86-0.97), showing that predicted answers capture the meaning well
4. **Overall Assessment:**
 - All predicted answers maintain high semantic similarity (>0.85), indicating strong meaning preservation
 - ROUGE-L scores suggest reasonable structural overlap

- Low BLEU scores are typical for short reference texts and don't indicate poor quality
- The evaluation shows that predictions successfully convey the correct information despite different phrasing

Conclusion

The evaluation demonstrates that the predicted answers effectively capture the semantic meaning of the actual answers, with an average semantic similarity of **0.9178**. While exact word-for-word matching (BLEU) is low, the high semantic similarity indicates that the predictions are semantically accurate and appropriate alternatives to the actual answers.