HybridRAG Evaluation Report
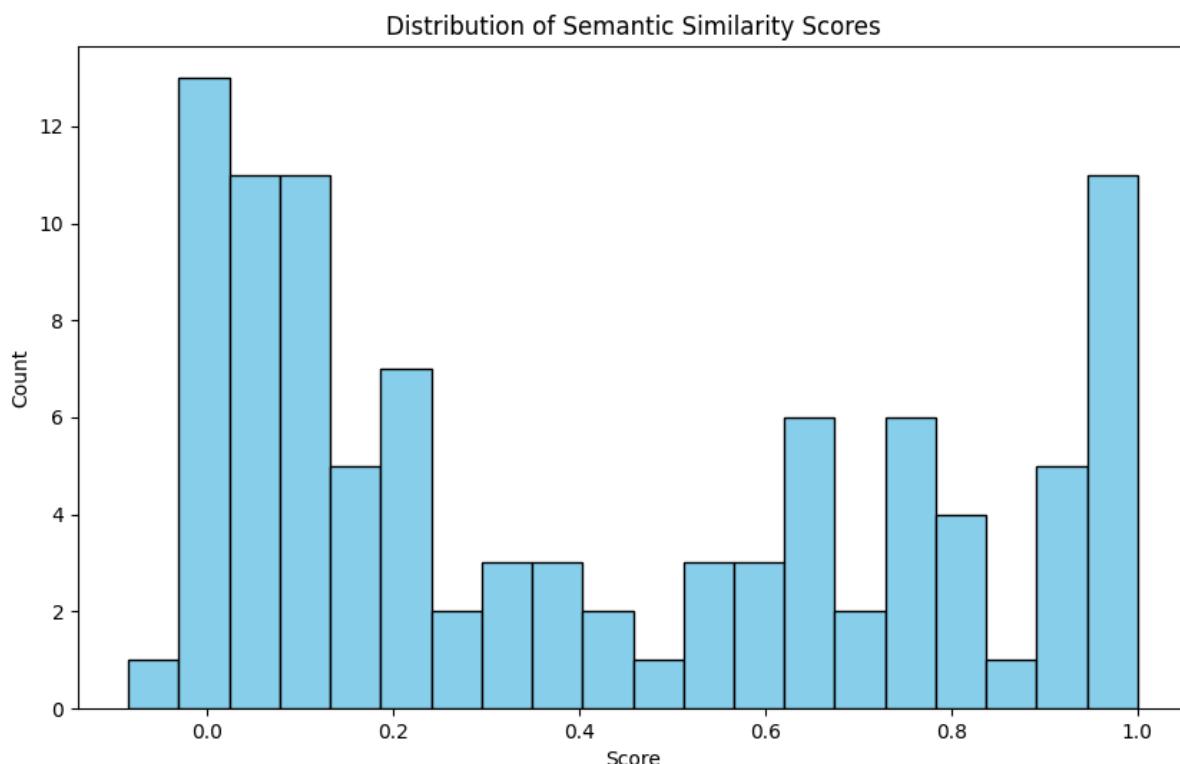
Generated automatically by the Evaluation Pipeline.

Executive Summary

| Metric	Score
Average MRR (Hybrid)	0.6831
Average Semantic Similarity	0.4035
Average BLEU Score	0.0932
Average Faithfulness (NLI)	0.7550
Average LLM Judge Score (1-5)	2.99
Average Response Time	0.98 s
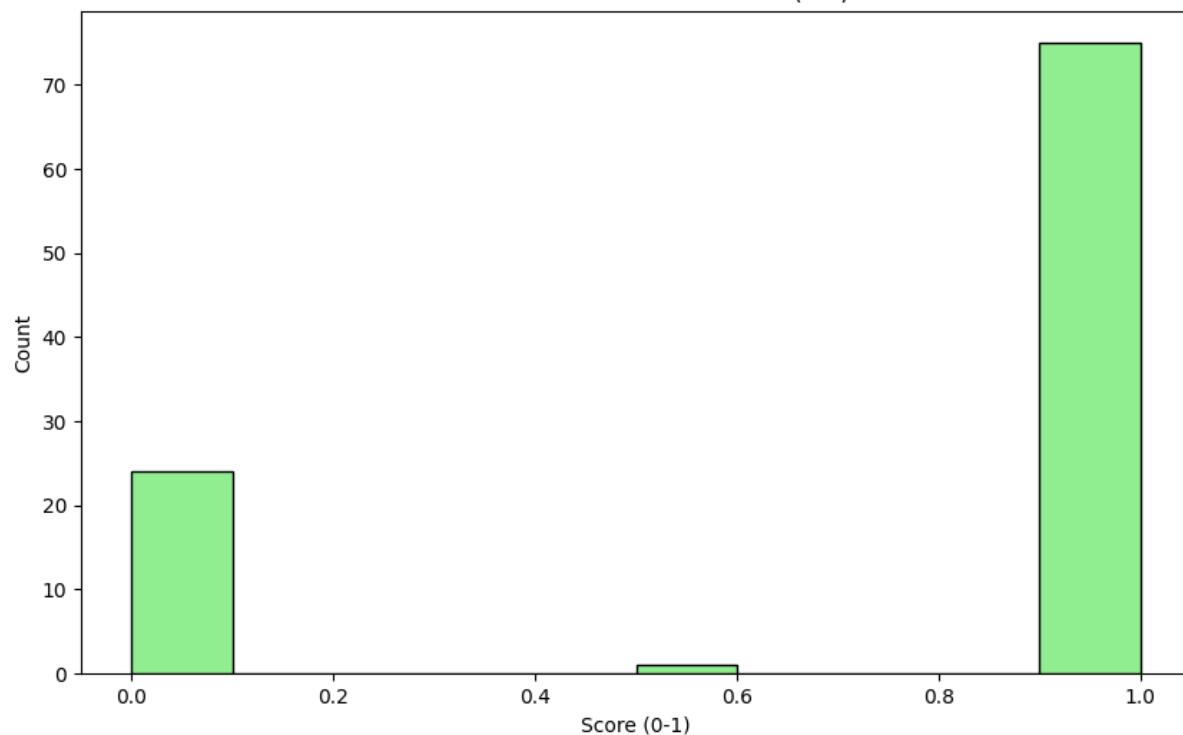
Visualizations

Metric Distributions

Distribution of Semantic Similarity Scores
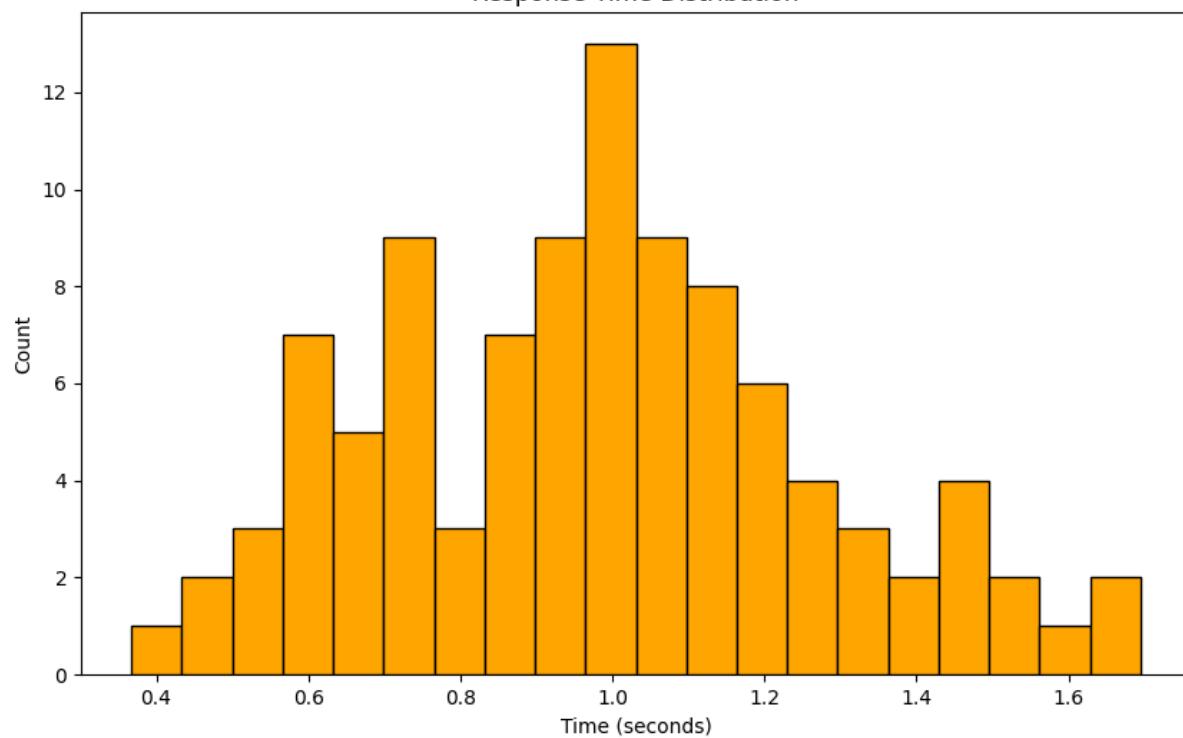
Faithfulness Distribution

Faithfulness Score Distribution (NLI)
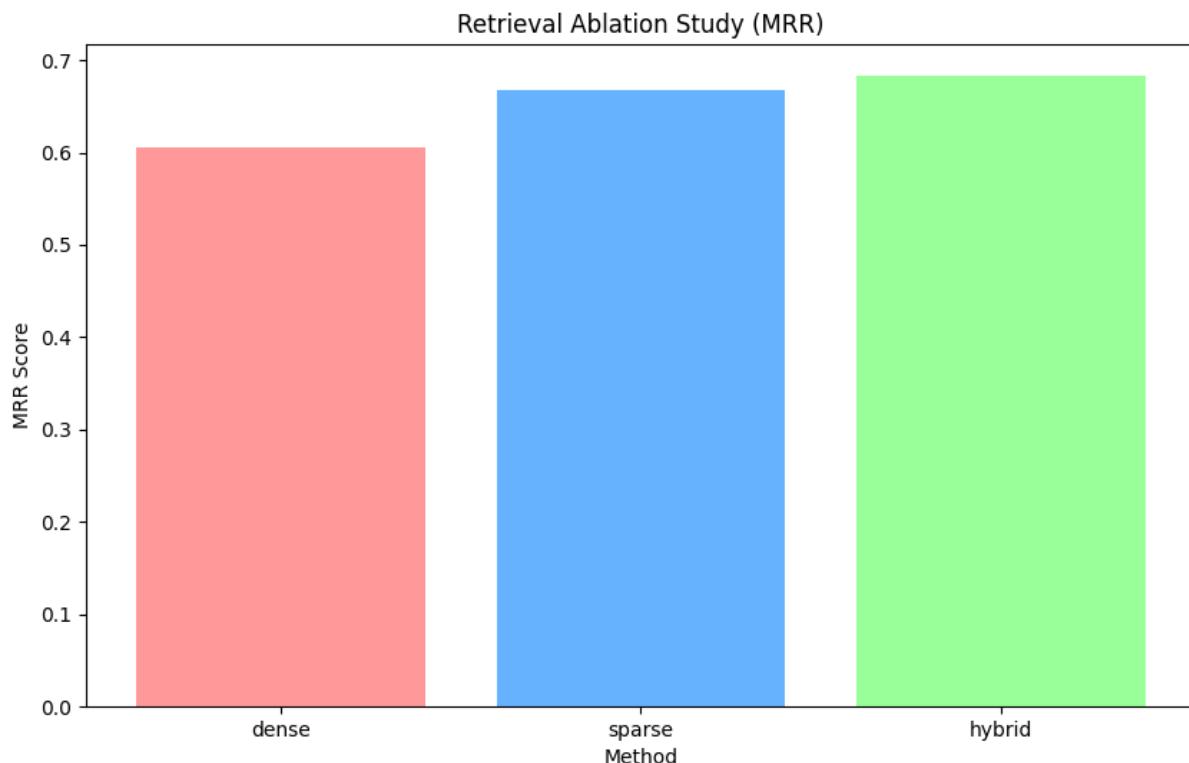
Response Time

Response Time Distribution

Retrieval Ablation

Retrieval Ablation Study (MRR)

Ablation Study Results

Method	MRR	Hit Rate
Dense	0.6059	0.7200
Sparse	0.6670	0.8000
Hybrid	0.6831	0.8300

Error Analysis (Bottom 5 Semantic Scores)

ID	Question	Semantic Score	Issue
77	What is the most common way to modify a CDS protocol?	-0.0837	Low Similarity
78	What is the term for a system that has been certified by an external security-auditing organization?	-0.0269	Low Similarity
65	What are the two systems that are interrelated?	-0.0148	Low Similarity
69	What is the private key used for?	-0.0115	Low Similarity

| 38 | What is the difference between the two GFs? | -0.0097 | Low Similarity |