HybridRAG Evaluation Report

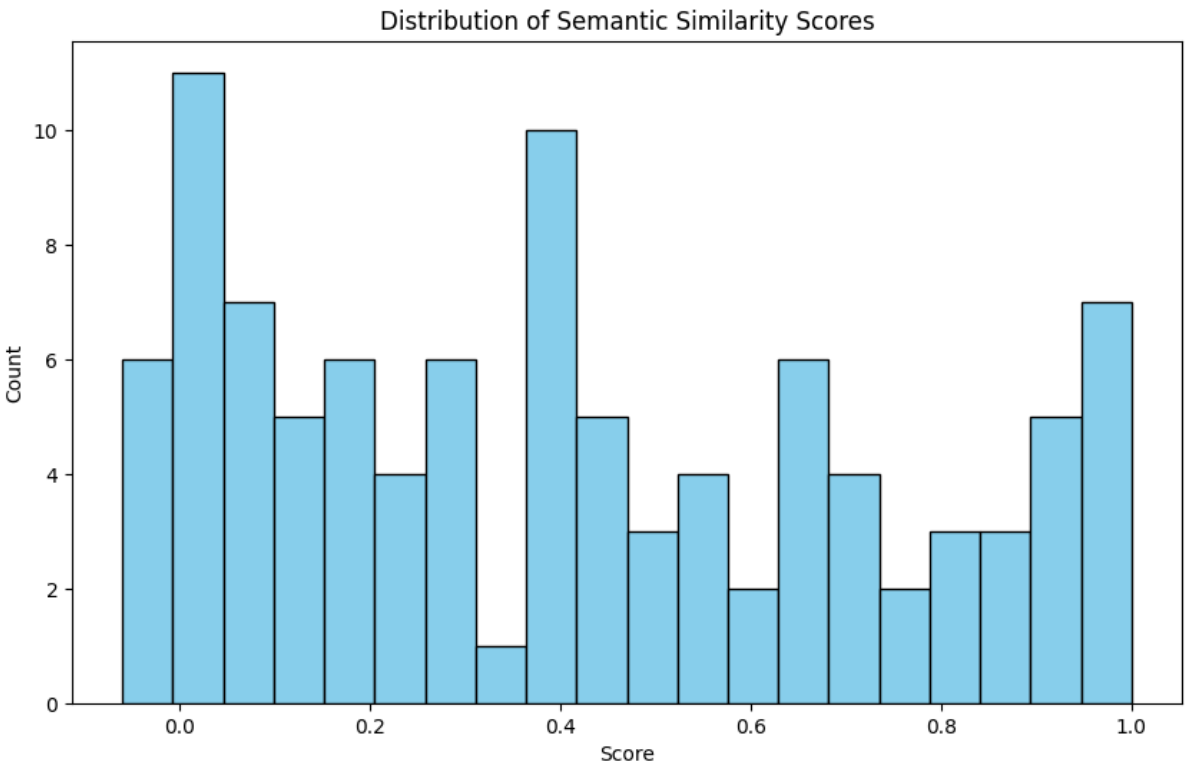
Generated automatically by the Evaluation Pipeline.

Executive Summary

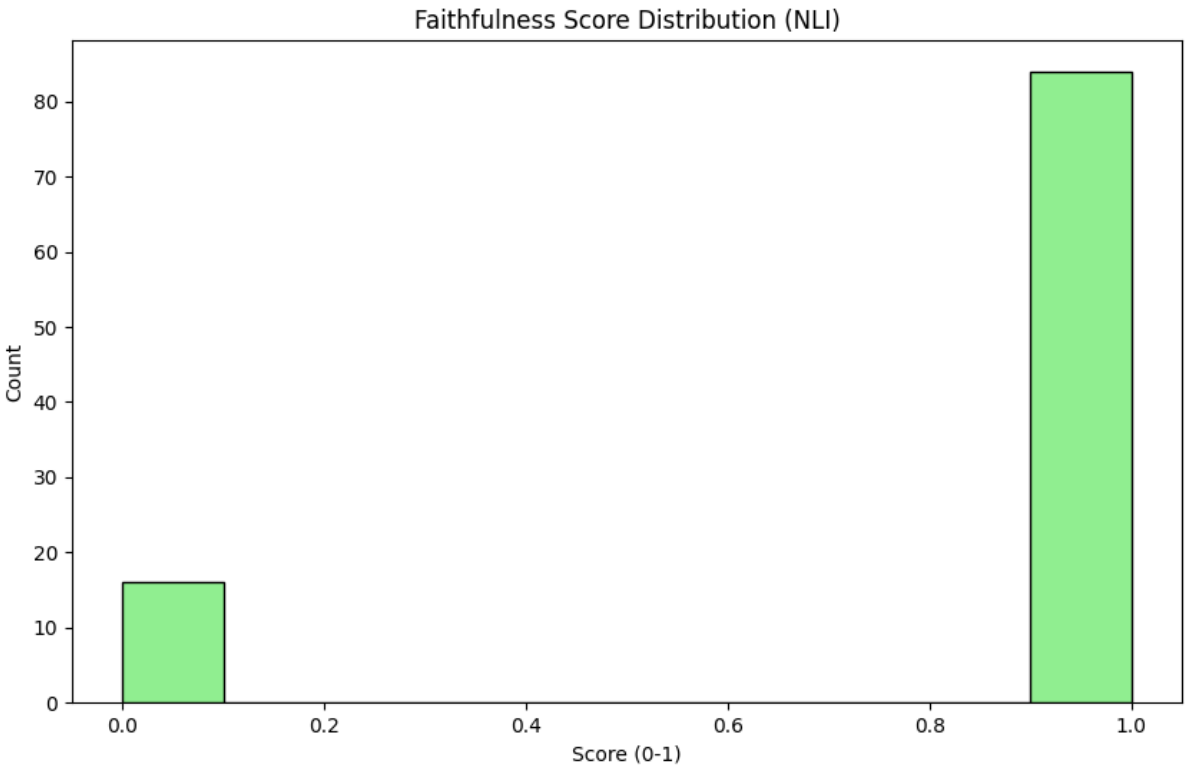
Metric	Score
Average MRR (Hybrid)	0.6653
Average Semantic Similarity	0.4095
Average BLEU Score	0.0851
Average Faithfulness (NLI)	0.8400
Average LLM Judge Score (1-5)	2.87
Average Response Time	1.09 s

Visualizations

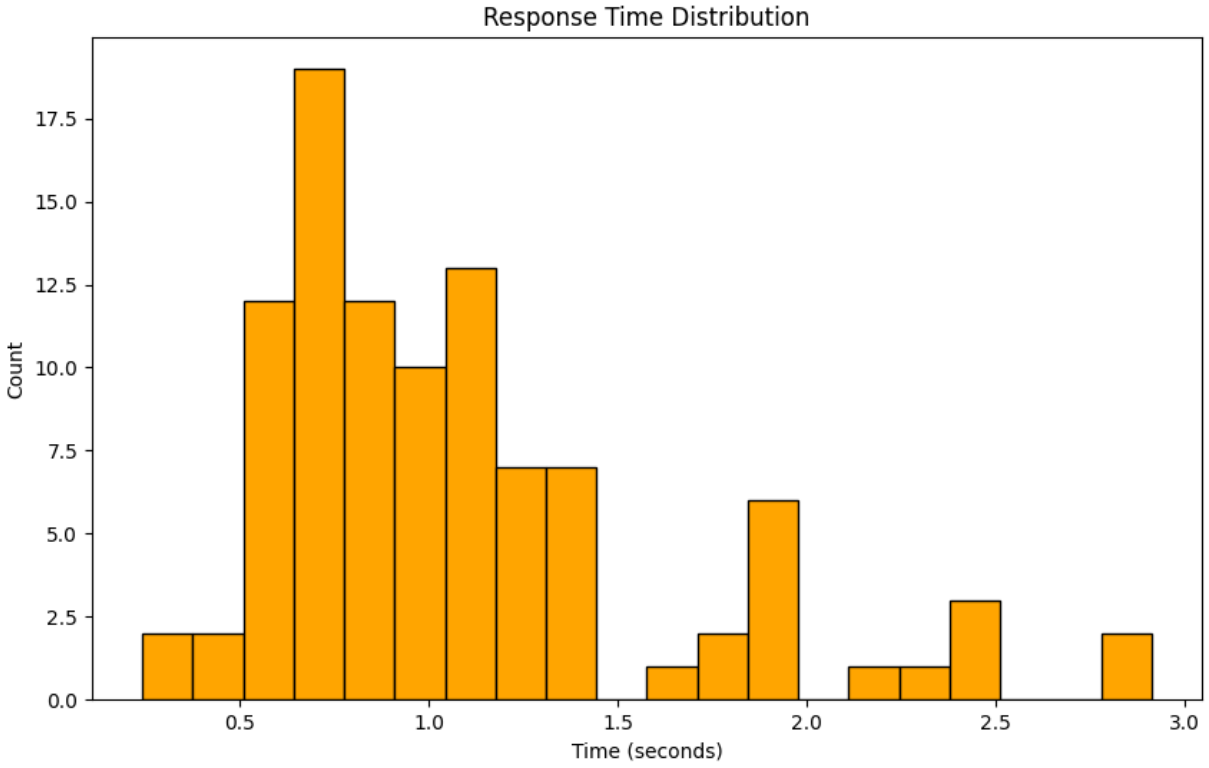
Metric Distributions



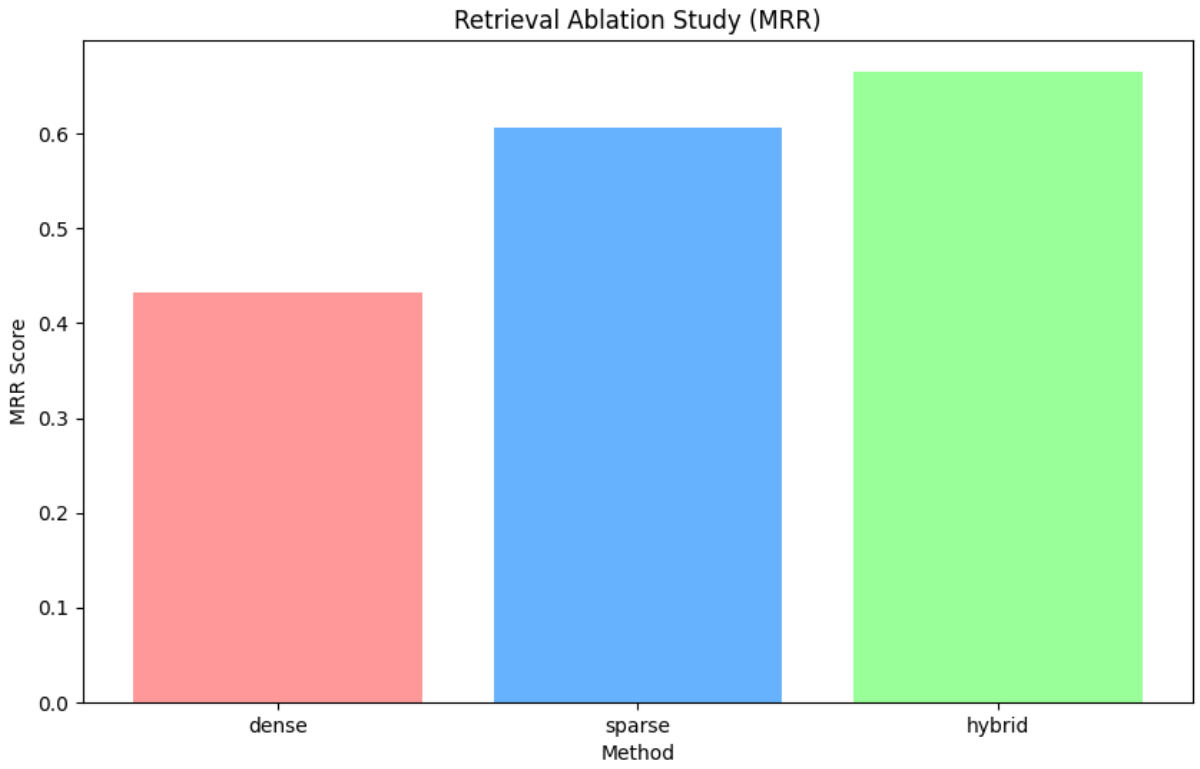
Faithfulness Distribution



Response Time



Retrieval Ablation



Ablation Study Results

Method	MRR	Hit Rate
Dense	0.4319	0.5900
Sparse	0.6059	0.6900
Hybrid	0.6653	0.7700

Error Analysis (Bottom 5 Semantic Scores)

ID	Question	Semantic Score	Issue
24	What is the name of the public key that Alice and Bob have?	-0.0597	Low Similarity
49	What is the difference between the two methods?	-0.0537	Low Similarity
46	What is the difference between a displaystyle and a displaystyle?	-0.0475	Low Similarity
48	What is the difference between SoK: Secure Messaging and SoK: Secure Messaging?	-0.0449	Low Similarity

73	What is the most important security vulnerability that HSTS can fix?	-0.0390	Low Similarity