Hybrid RAG System Evaluation Report

1. System Architecture

User Query -> [Dense Retrieval (FAISS)] + [Sparse Retrieval (BM25)] -> RRF Fusion -> Context Chunks -> LLM Generation (Flan-T5) -> Answer

Data Flow:
- URL Collection -> Text Extraction -> Chunking -> Indexing (Dense/Sparse)
- Evaluation: Q&A Generation -> Query Processing -> Metrics (MRR, Similarity, etc.) -> Reports

2. Evaluation Results

Metric	Value
MRR	0.0000
Semantic_Similarity	0.67714286
Answer_Relevance	0.2806057

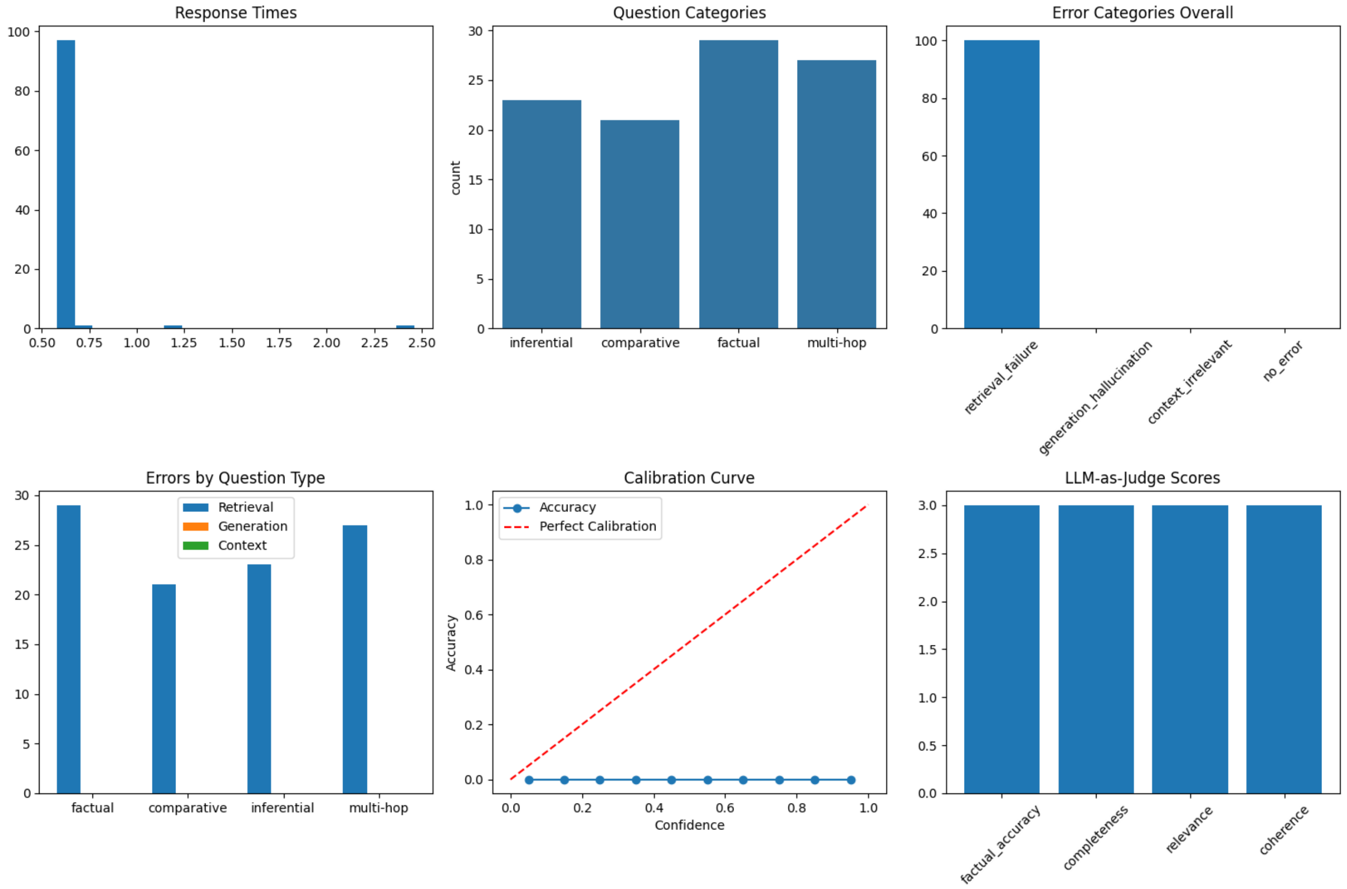
Error Analysis

Retrieval Failure	100
Generation Hallucination	0
Context Irrelevant	0
No Error	0

3. Innovative Evaluation Approaches

- Error Analysis: Automatic categorization of failures (retrieval, generation, context) by question type.
- LLM-as-Judge: Uses Flan-T5 to evaluate answers on factual accuracy, completeness, relevance, and coherence.
- Confidence Calibration: Measures calibration with Expected Calibration Error (ECE) and provides calibration curves.
- Ablation Studies: Compare dense-only, sparse-only, and hybrid retrieval performance (not implemented in this run).

4. Visualizations



Sample Query Results

Query	Answer	MRR	Similarity
What does soil liquefacti...	when, because of the shaking, water-satu...	0.00	0.68
What is the text mainly a...	A...	0.00	0.68
What type of art is being...	Hindu sculpture...	0.00	0.68
What's the main idea of t...	C...	0.00	0.68
What did the Vijayanagara...	conquered the entire Tamil country...	0.00	0.68