RAG Evaluation Report

Overall Performance Summary

mr_url: 0.9683
precision_at_5_url: 0.2280
answer_f1: 0.0517
contextual_precision: 0.9853
distinct_1: 0.5829
avg_latency: 2.2639

Custom Metrics Justification

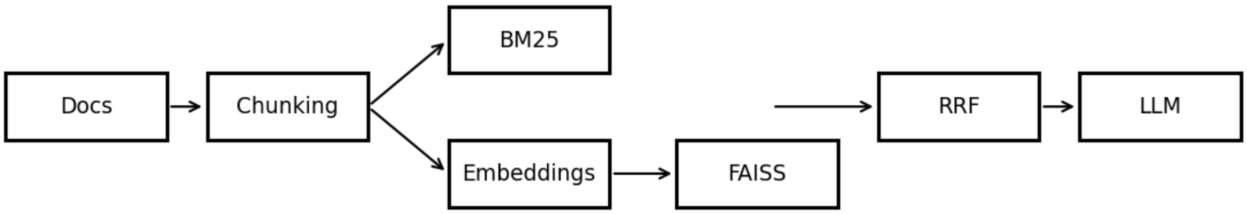
Precision@5 (URL-level)

Why: retrieval quality in the small window that feeds the generator.
How: count correct URLs in top-5 unique URLs, divide by 5, average.
Interpretation: higher is better; 0.6 ≈ 3/5 URLs relevant.

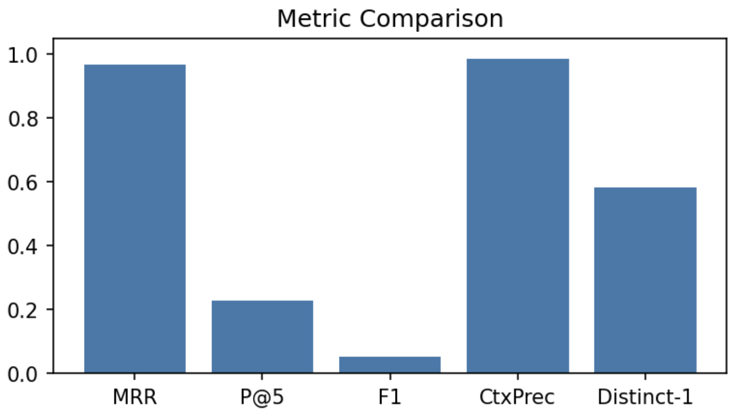
Answer F1 (token overlap)

Why: captures answer quality with partial credit.
How: tokenize; precision/recall on overlap; F1 = 2PR/(P+R).
Interpretation: higher is better; 1.0 = perfect lexical match.
Answer F1: precision/recall overlap; F1 = 2PR/(P+R).

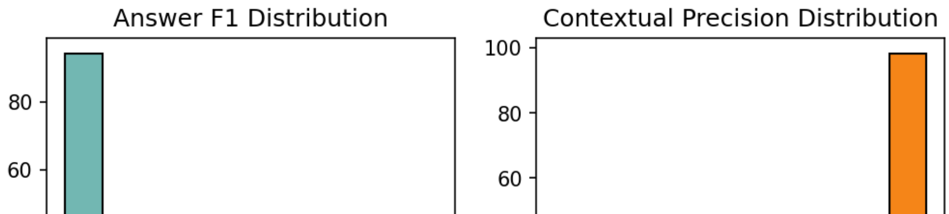
Architecture Diagram

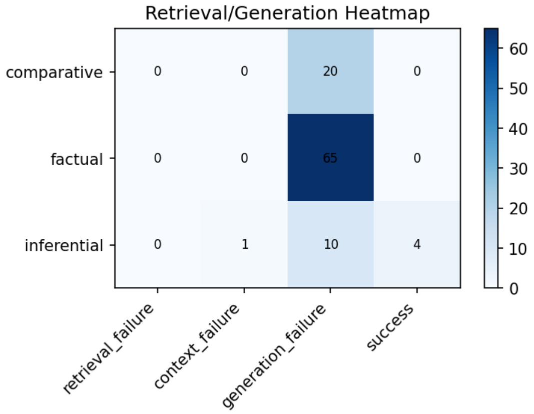


Metric Comparison

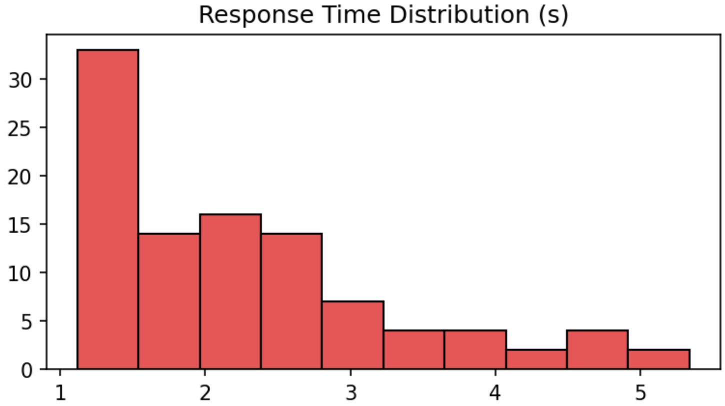


Score Distributions

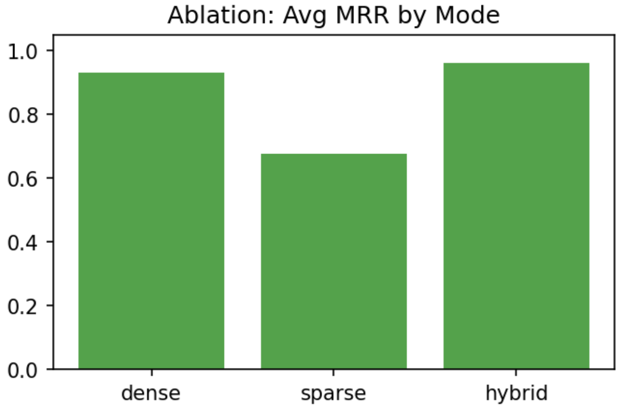




Response Times



Ablation Plot



Adversarial Testing

Innovative study for adversarial testing:

We stress-test retrieval and generation with paraphrases and unanswerable variants.
Paraphrase tests check robustness of URL ranking; unanswerables probe hallucination risk.
A response is flagged hallucinated if it is non-empty but poorly grounded in context.
paraphrase_hit_rate: 0.9000
unanswerable_hallucination_rate: 0.1000

Error Analysis (by category)

factual: {'generation_failure': 65}

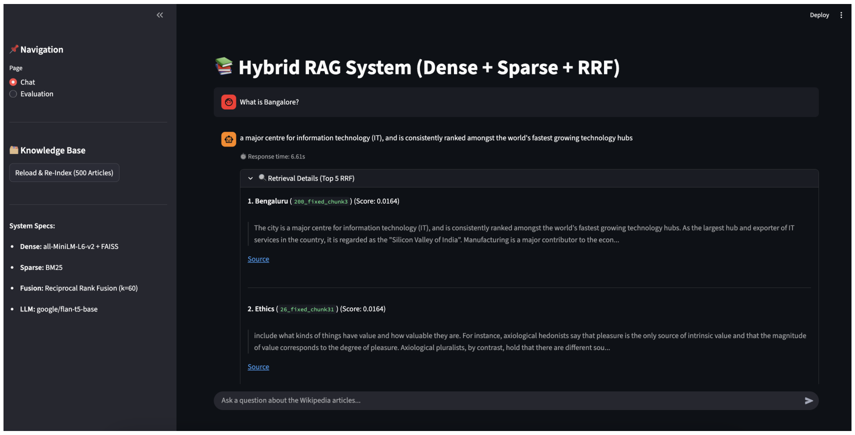
comparative: {'generation_failure': 20}
inferential: {'generation_failure': 10, 'success': 4, 'context_failure': 1}

Pattern: Most failures by category: factual (65).

Results Table (first 10, includes MRR contribution)

1: MRR=1.00 F1=0.00 Ctx=0.93 T=4.70s
2: MRR=1.00 F1=0.03 Ctx=1.00 T=2.01s
3: MRR=1.00 F1=0.00 Ctx=1.00 T=2.29s
4: MRR=1.00 F1=0.00 Ctx=1.00 T=1.44s
5: MRR=1.00 F1=0.00 Ctx=1.00 T=1.18s
6: MRR=1.00 F1=0.00 Ctx=1.00 T=1.41s
7: MRR=1.00 F1=0.00 Ctx=1.00 T=2.26s
8: MRR=1.00 F1=0.02 Ctx=1.00 T=2.43s
9: MRR=1.00 F1=0.00 Ctx=1.00 T=2.57s
10: MRR=1.00 F1=0.00 Ctx=1.00 T=2.93s

Screenshots



System Specs:

- Dense: all-MiniLM-L6-v2 + FAISS
- Sparse: BM25
- Fusion: Reciprocal Rank Fusion (k=60)
- LLM: google/flan-t5-base

Hybrid RAG System (Dense + Sparse + RRF)

What is Bangalore?

a major centre for information technology (IT), and is consistently ranked amongst the world's fastest growing technology hubs

Response time: 6.61s

Retrieval Details (Top 5 RRF)

1. Bengaluru (`200_fixed_chunk3`) (Score: 0.0164)

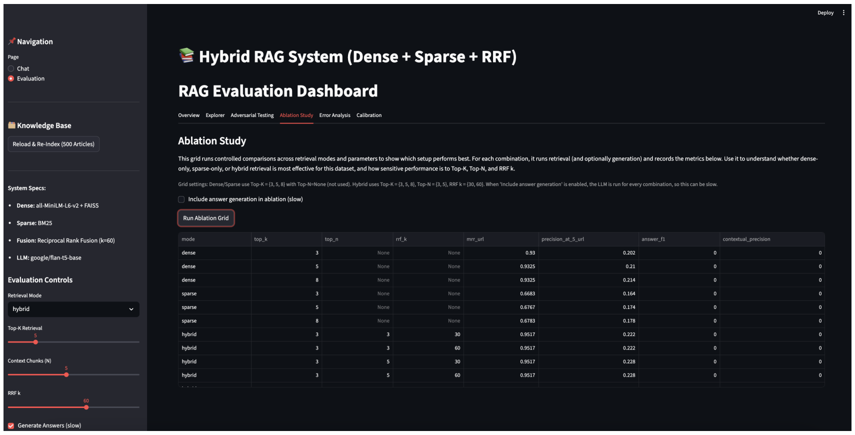
> The city is a major centre for information technology (IT), and is consistently ranked amongst the world's fastest growing technology hubs. As the largest hub and exporter of IT services in the country, it is regarded as the "Silicon Valley of India". Manufacturing is a major contributor to the econ...

Source

2. Ethics (`26_fixed_chunk31`) (Score: 0.0164)

> include what kinds of things have value and how valuable they are. For instance, axiological hedonists say that pleasure is the only source of intrinsic value and that the magnitude of value corresponds to the degree of pleasure. Axiological pluralists, by contrast, hold that there are different sou...

Source



System Specs:

- Dense: all-MiniLM-L6-v2 + FAISS
- Sparse: BM25
- Fusion: Reciprocal Rank Fusion (k=60)
- LLM: google/flan-t5-base

Evaluation Controls

Retrieval Mode: hybrid

Top-K Retrieval: 5

Context Chunks (N): 5

RRF k: 60

☑ Generate Answers (slow)

Hybrid RAG System (Dense + Sparse + RRF)

RAG Evaluation Dashboard

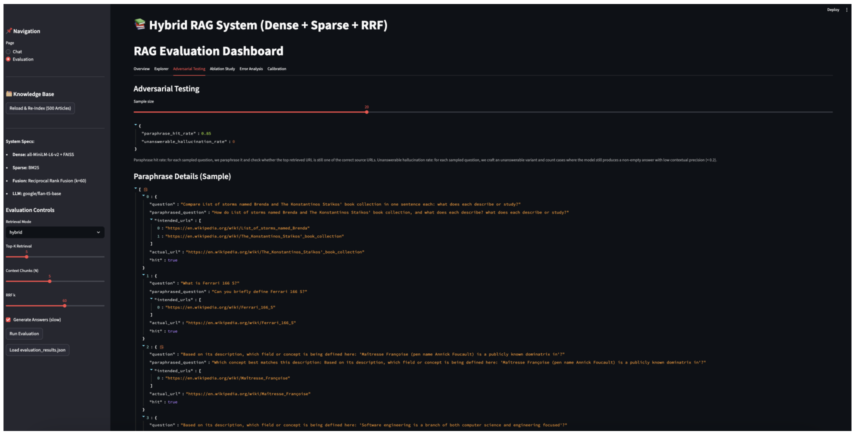
Ablation Study

This grid runs controlled comparisons across retrieval modes and parameters to show which setup performs best. For each combination, it runs retrieval (and optionally generation) and records the metrics below. Use it to understand whether dense-only, sparse-only, or hybrid retrieval is most effective for this dataset, and how sensitive performance is to Top-K, Top-N, and RRF k.

Grid settings: Dense/Sparse use Top-K = {3, 5, 8} with Top-N=None (not used). Hybrid uses Top-K = {3, 5, 8}, Top-N = {3, 5}, RRF k = {30, 60}. When 'Include answer generation' is enabled, the LLM is run for every combination, so this can be slow.

☐ Include answer generation in ablation (slow)

mode	top_k	top_n	rrf_k	mrr_url	precision_at_5_url	answer_f1	contextual_precision
dense	3	None	None	0.93	0.202	0	0
dense	5	None	None	0.9325	0.21	0	0
dense	8	None	None	0.9325	0.214	0	0
sparse	3	None	None	0.6683	0.164	0	0
sparse	5	None	None	0.6767	0.174	0	0
sparse	8	None	None	0.6783	0.178	0	0
hybrid	3	3	30	0.9517	0.222	0	0
hybrid	3	3	60	0.9517	0.222	0	0
hybrid	3	5	30	0.9517	0.228	0	0
hybrid	3	5	60	0.9517	0.228	0	0



System Specs:

- Dense: all-MiniLM-L6-v2 + FAISS
- Sparse: BM25
- Fusion: Reciprocal Rank Fusion (k=60)
- LLM: google/flan-t5-base

Evaluation Controls

Retrieval Mode: hybrid

Top-K Retrieval: 5

Context Chunks (N): 5

RRF k: 60

☑ Generate Answers (slow)

Hybrid RAG System (Dense + Sparse + RRF)

RAG Evaluation Dashboard

Adversarial Testing

Sample size: 20

```
{
  "paraphrase_hit_rate" : 0.85
  "unanswerable_hallucination_rate" : 0
}
```

Paraphrase hit rate: for each sampled question, we paraphrase it and check whether the top retrieved URL is still one of the correct source URLs. Unanswerable hallucination rate: for each sampled question, we craft an unanswerable variant and count cases where the model still produces a non-empty answer with low contextual precision (< 0.2).

Paraphrase Details (Sample)

```
[
  0 : {
    "question" : "Compare List of storms named Brenda and The Konstantinos Staikos' book collection in one sentence each: what does each describe or study?"
    "paraphrased_question" : "How do List of storms named Brenda and The Konstantinos Staikos' book collection, and what does each describe? what does each describe or study?"
    "intended_urls" : [
      0 : "https://en.wikipedia.org/wiki/List_of_storms_named_Brenda"
      1 : "https://en.wikipedia.org/wiki/The_Konstantinos_Staikos'_book_collection"
    ]
    "actual_url" : "https://en.wikipedia.org/wiki/The_Konstantinos_Staikos'_book_collection"
    "hit" : true
  }
  1 : {
    "question" : "What is Ferrari 166 S?"
    "paraphrased_question" : "Can you briefly define Ferrari 166 S?"
    "intended_urls" : [
      0 : "https://en.wikipedia.org/wiki/Ferrari_166_S"
    ]
    "actual_url" : "https://en.wikipedia.org/wiki/Ferrari_166_S"
    "hit" : true
  }
  2 : {
    "question" : "Based on its description, which field or concept is being defined here: 'Maitresse Françoise (pen name Annick Foucault) is a publicly known dominatrix in'?"
    "paraphrased_question" : "Which concept best matches this description: Based on its description, which field or concept is being defined here: 'Maitresse Françoise (pen name Annick Foucault) is a publicly known dominatrix in'?"
    "intended_urls" : [
      0 : "https://en.wikipedia.org/wiki/Maitresse_Françoise"
    ]
    "actual_url" : "https://en.wikipedia.org/wiki/Maitresse_Françoise"
    "hit" : true
  }
  3 : {
    "question" : "Based on its description, which field or concept is being defined here: 'Software engineering is a branch of both computer science and engineering focused'?"
```