# RAG Evaluation Report

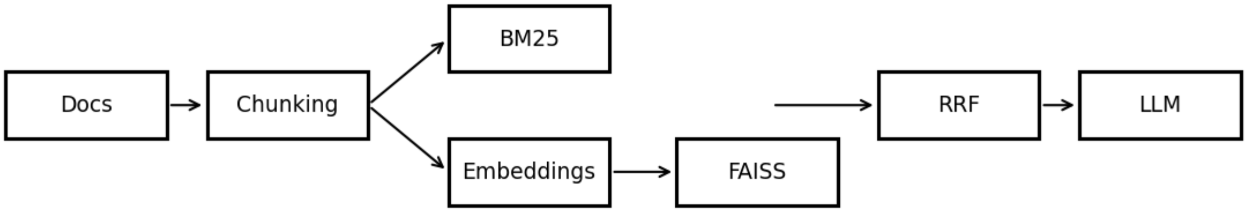
## Overall Performance Summary

mrr_url: 0.9683
precision_at_5_url: 0.2280
answer_f1: 0.0517
contextual_precision: 0.9853
distinct_1: 0.5829
avg_latency: 2.0042

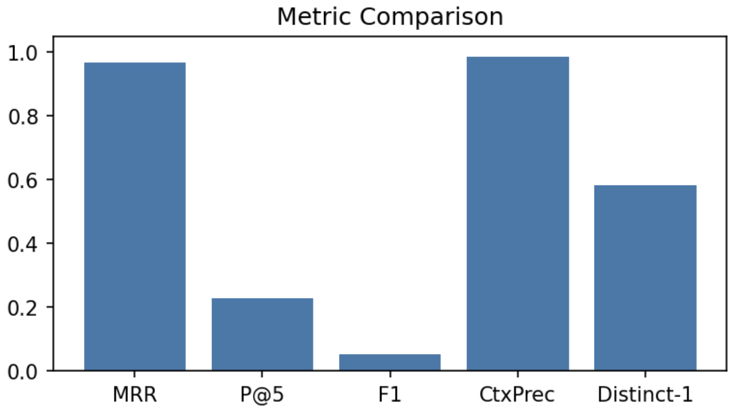
## Custom Metrics Justification

Answer F1: precision/recall overlap; F1 = 2PR/(P+R).
Contextual Precision: |answer tokens ∩ context tokens| / |answer tokens|.

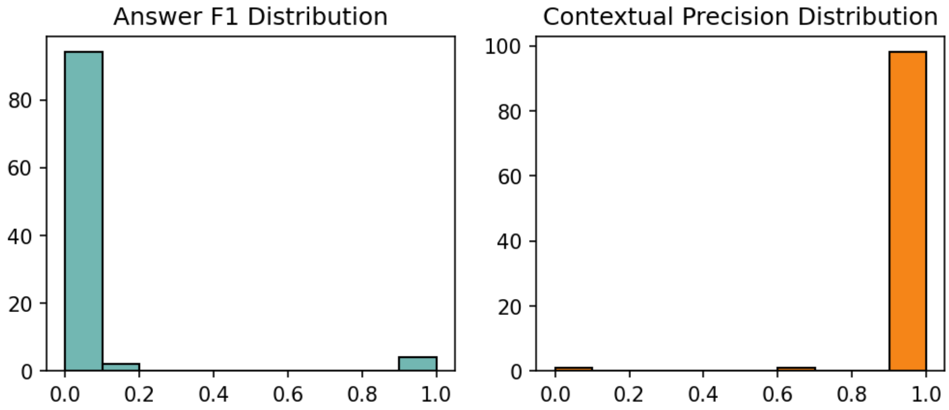
## Architecture Diagram



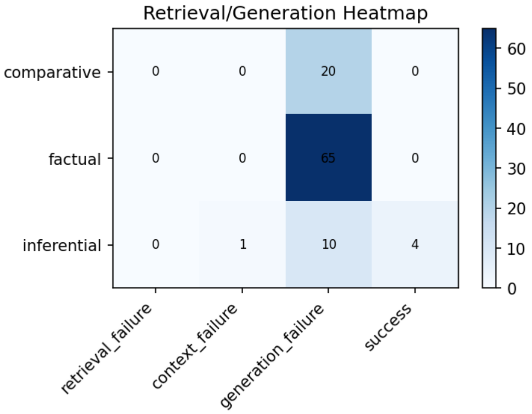
## Metric Comparison



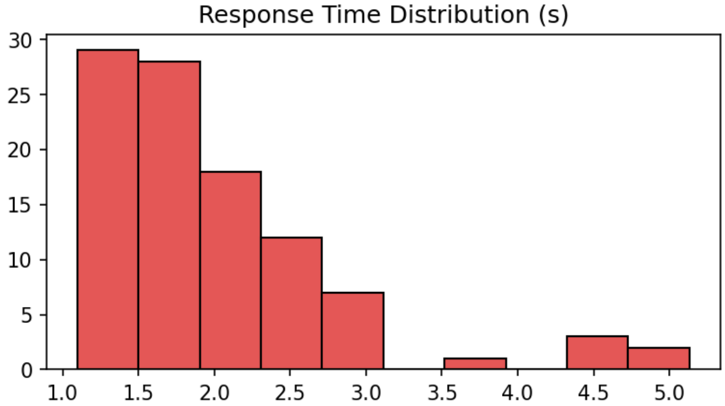
## Score Distributions



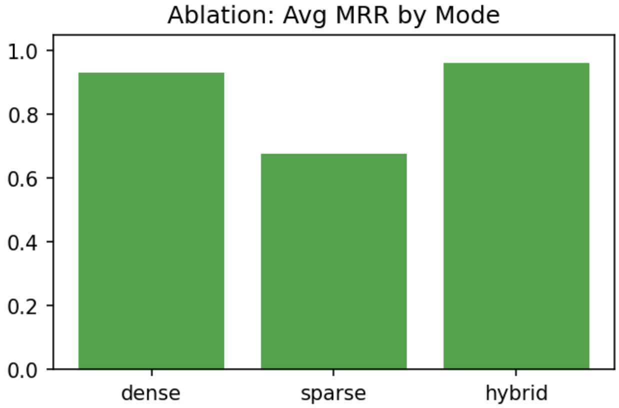
## Retrieval Heatmap



### Response Times



### Ablation Plot



### Adversarial Testing

paraphrase_hit_rate: 0.9500<br>
unanswerable_hallucination_rate: 0.0000

### 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=4.58s
2: MRR=1.00 F1=0.03 Ctx=1.00 T=2.62s
3: MRR=1.00 F1=0.00 Ctx=1.00 T=2.79s
4: MRR=1.00 F1=0.00 Ctx=1.00 T=1.97s
5: MRR=1.00 F1=0.00 Ctx=1.00 T=1.38s
6: MRR=1.00 F1=0.00 Ctx=1.00 T=1.32s
7: MRR=1.00 F1=0.00 Ctx=1.00 T=2.67s
8: MRR=1.00 F1=0.02 Ctx=1.00 T=2.69s
9: MRR=1.00 F1=0.00 Ctx=1.00 T=2.25s
10: MRR=1.00 F1=0.00 Ctx=1.00 T=2.27s

**Screenshots**

Screenshot Placeholder 1

Screenshot Placeholder 2

Screenshot Placeholder 3