### RAG Evaluation Report

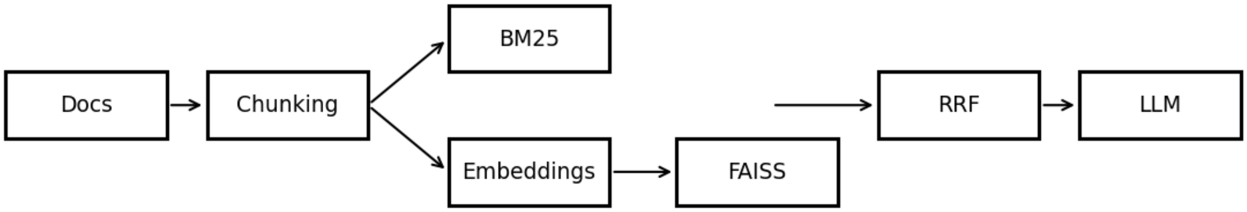
### Overall Performance Summary

mrr_url: 0.9417
precision_at_5_url: 0.2640
answer_f1: 0.3430
contextual_precision: 0.9900
distinct_1: 0.5241
avg_latency: 1.7913

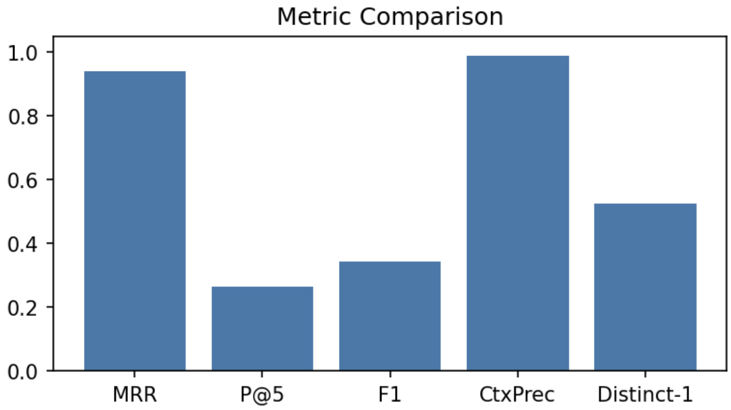
### Custom Metrics Justification

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

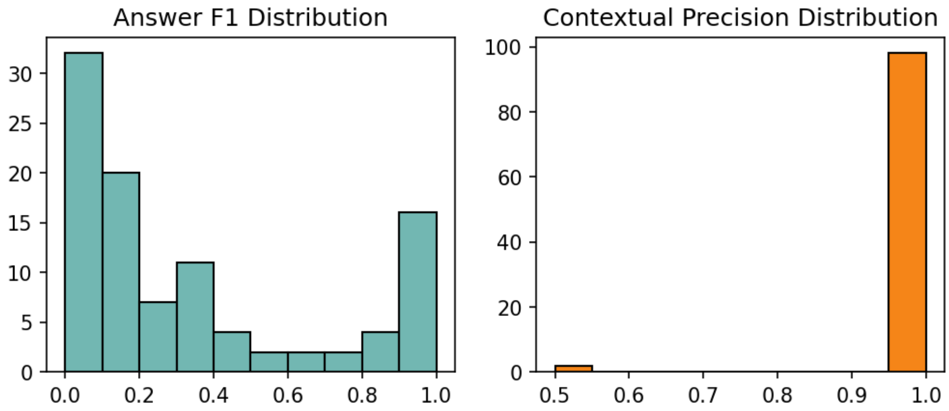
### Architecture Diagram



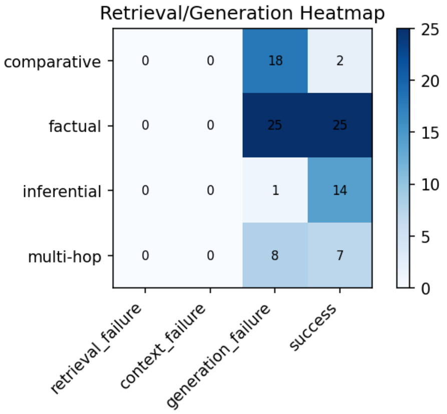
### Metric Comparison



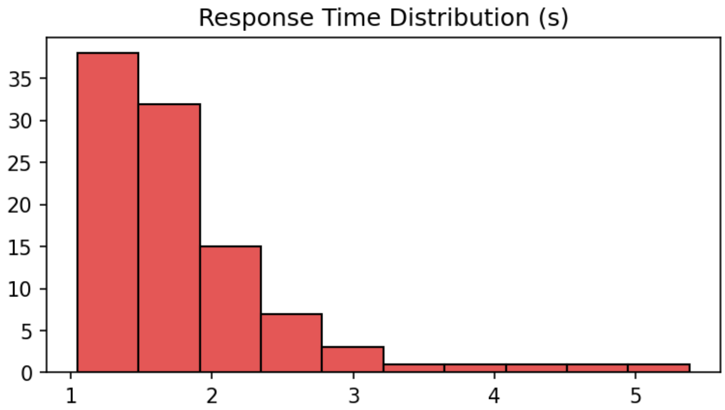
### Score Distributions



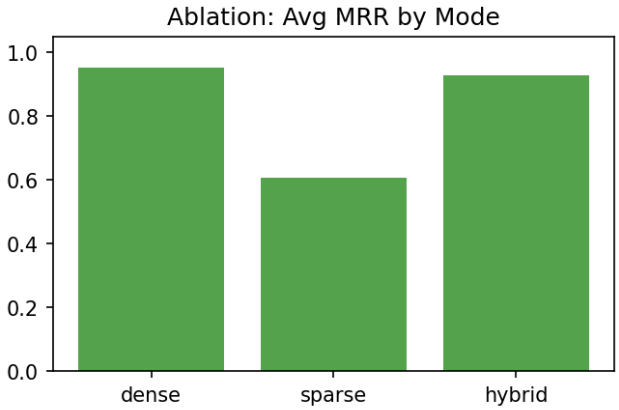
### Retrieval Heatmap



## Response Times



## Ablation Plot



## Adversarial Testing

paraphrase_hit_rate: 0.9500
unanswerable_hallucination_rate: 0.0500

## Error Analysis (by category)

factual: {'success': 25, 'generation_failure': 25}
comparative: {'generation_failure': 18, 'success': 2}
inferential: {'success': 14, 'generation_failure': 1}
multi-hop: {'generation_failure': 8, 'success': 7}

Pattern: Most failures by category: factual (25).

**Results Table (first 10, includes MRR contribution)**

```
1: MRR=1.00 F1=0.38 Ctx=1.00 T=3.49s
2: MRR=1.00 F1=0.08 Ctx=1.00 T=1.96s
3: MRR=1.00 F1=0.27 Ctx=1.00 T=1.70s
4: MRR=1.00 F1=0.11 Ctx=1.00 T=1.59s
5: MRR=1.00 F1=0.10 Ctx=1.00 T=1.71s
6: MRR=1.00 F1=0.00 Ctx=1.00 T=2.39s
7: MRR=1.00 F1=0.00 Ctx=1.00 T=1.38s
8: MRR=0.50 F1=0.88 Ctx=1.00 T=1.82s
9: MRR=1.00 F1=0.40 Ctx=1.00 T=1.94s
10: MRR=1.00 F1=0.48 Ctx=1.00 T=1.43s
```

**Screenshots**

Screenshot Placeholder 1

Screenshot Placeholder 2

Screenshot Placeholder 3