RAG retrieval evaluation

TECHNICAL REVIEW / 200 SYNTHETIC QUERIES

Retrieval quality

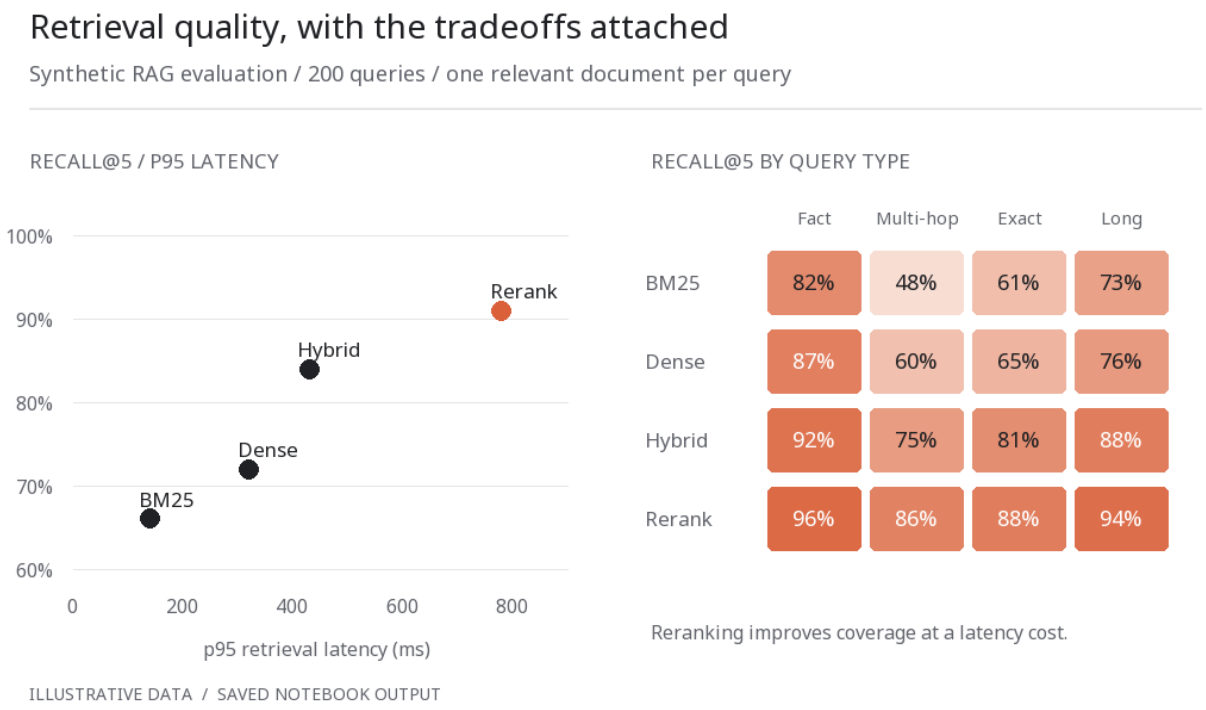
With one relevant document per query, Recall@5 is the share retrieving it in the top five. Compare quality and latency together.

```
runs = [["BM25", .66, .52, 140], ["Dense", .72, .58, 320],
        ["Hybrid", .84, .71, 430], ["Rerank", .91, .83, 780]]
import pandas as pd
results = pd.DataFrame(runs, columns=["Retriever", "Recall@5", "MRR", "p95 ms"])
results
```

Retriever	Recall@5	MRR	p95 ms
BM25	0.66	0.52	140
Dense	0.72	0.58	320
Hybrid	0.84	0.71	430
Rerank	0.91	0.83	780

Reranking reaches 91% Recall@5 at 780 ms p95; hybrid search reaches 84% at 430 ms. Query slices below are illustrative and do not reconstruct the overall score.

```
import matplotlib.pyplot as plt
slice_recall = [[.82, .48, .61, .73], [.87, .60, .65, .76],
                [.92, .75, .81, .88], [.96, .86, .88, .94]]
fig, axes = plt.subplots(1, 2, figsize=(11, 5))
axes[0].scatter(results["p95 ms"], results["Recall@5"])
axes[0].set(xlabel="p95 ms", ylabel="Recall@5")
axes[1].imshow(slice_recall, cmap="Oranges", vmin=.4, vmax=1)
axes[1].set_yticks(range(4), results["Retriever"])
axes[1].set_xticks(range(4), ["Fact", "Multi-hop", "Exact", "Long"])
plt.show()
```



Review the evaluation, not just the score

$$MRR = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{rank_q}$$

A missing relevant document contributes zero. Validate query representativeness, relevance judgments, latency percentiles and leakage before choosing a retriever. These invented values are not measurements of any real service.

Full Notebook preserves the source, evidence and caveats for a technical reviewer.