Queryable Shared Reference Repository

Building an intelligent, privacy-preserving system for scientific research lab (VITEK)

Atyab Hakeem, Kishan Sathish Babu, Naga Kushal Ageeru, Pranav Kanth Anbarasan

Objective

Growing Volume

Research groups struggle to manage ever-increasing scientific literature.

Limited Search

Current reference managers lack intelligent, context-aware querying capabilities

Privacy Concerns

Cloud-based LLM's raise data privacy issues and produce hallucinated outputs

Data Source

Sources & Formats:

- **Scientific Papers**:
  - PDF's and webpages
  - Variable layouts (journal / publisher differences)

- **Metadata Files:**
  - Formats: .bib (BibTeX) and reference documents
  - Enable citation and filtering

Documents	297
Avg Words	1,782
Vocabulary	39,144
Avg Tables	1
Avg Figures	7

Volume and Scale

Current Capacity

- 300 scientific paper

Scalability:

- Expandable to 10,000 papers

User Access:

- 1 - 3 concurrent users
- Max 10 lab members

Solution - Retrieval Augmented Generation

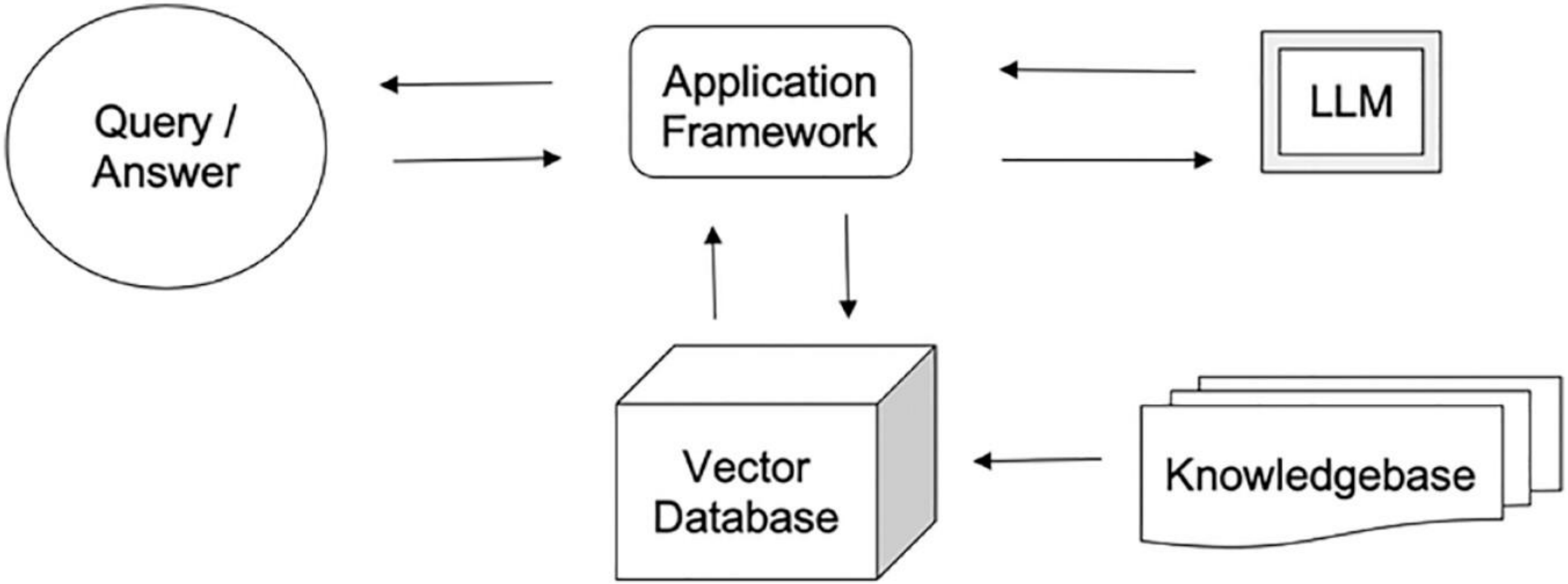
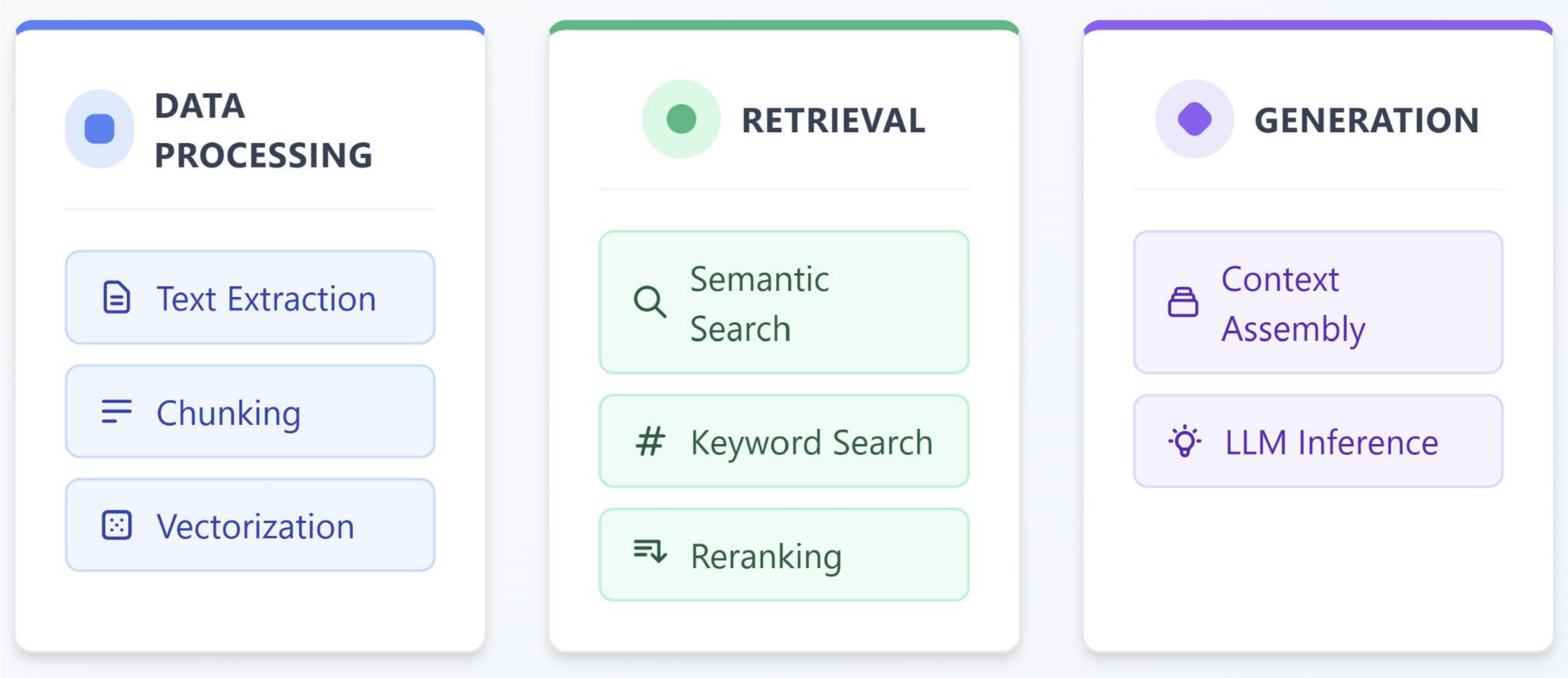


Figure 1: A simple Retrieval Augmented Generation (RAG) system

Mashatian et al. (2024). Building Trustworthy Generative AI for Diabetes Care. J Diabetes Sci Technol, 19(5):1264-1270.

Retrieval Augmented Generation Components



Data Processing - Parsing, Chunking, & Embedding

Parsing

- Extracting text from documents

Chunking

- Splitting text into smaller segments
- Improves
  - Precision
  - Information captured
  - Response Quality
- Recursive splitter uses priority

Embedding

- Conversion to vectors
- Similarity search to fetch similar text

```
The degree to which the returns for
performance are superlinear.
```

Character Splitter; Chunk size = 25; Overlap = 0

```
The degree to which the returns for
performance are superlinear.
```

Character Splitter; Chunk size = 10; Overlap = 3

```
The degree to which the r ➔ [[0.5..]..]
eturns for performance ar ➔ [[0.3..]..]
e superlinear. ➔ [[0.6..]..]
```

Source: chunkviz.up.railway.app

Hit Rate

Hit Rate = $1/n \sum_{i=1}^{n} \mathbb{1}(\text{recall}_i > \tau)$

$\text{recall}_i = |D_i \cap G_i| / |G_i|$

n = number of queries

D_i = retrieved documents for query i

G_i = ground truth documents for query i

τ = threshold

$\mathbb{1}(\cdot)$ = indicator function

Query: What is the typical outcome of a MALDI-imaging study?

Ground Truth:
A typical MALDI-imaging study results in a set of ions of interest

Retrieved Documents:
MALDI-imaging study results in a set of ions of interest. confocal microscopy imaging techniques. experimental outcomes vary significantly.

$\tau = 0.5$	$\tau = 0.9$
$\|D_i \cap G_i\| = 11$	$\|D_i \cap G_i\| = 11$
$\|G_i\| = 13$	$\|G_i\| = 13$
Recall = 0.846	Recall = 0.846
Hit Rate = 1	**Hit Rate = 0**

Mean Reciprocal Rank

$$\mathbf{MRR = 1/n \sum_{i=1}^{n} 1/rank_i}$$

$rank_i$ = rank of the first relevant document

relevant document = $recall_i > \tau$

$recall_i = |D_i \cap G_i| / |G_i|$

τ = threshold

Query: What is the typical outcome of a MALDI-imaging study?

Ground Truth:
A typical MALDI-imaging study results in a set of ions of interest

Retrieved Documents:
1. confocal microscopy imaging techniques.
2. MALDI-imaging study results in a set of ions of interest
3. experimental outcomes vary significantly.

$\tau = 0.5$	$\tau = 0.9$
$\|D_i \cap G_i\| = 11$; $\|G_i\| = 13$	$\|D_i \cap G_i\| = 11$; $\|G_i\| = 13$
Rank = 2	Rank = 2
Recall = 0.846	Recall = 0.846
MRR = 0.5	**MRR = 0**

Phase 1 Data Processing Limitations

Issues

- Limited Performance
- Fragmented Chunks
- Missing Words/ Unknowns

Solution

- Larger Models
- Non-OCR, Non-LLM Vision Model augmented data processor (Docling)

Docling + Gemma Provide Top Scores & Text Quality

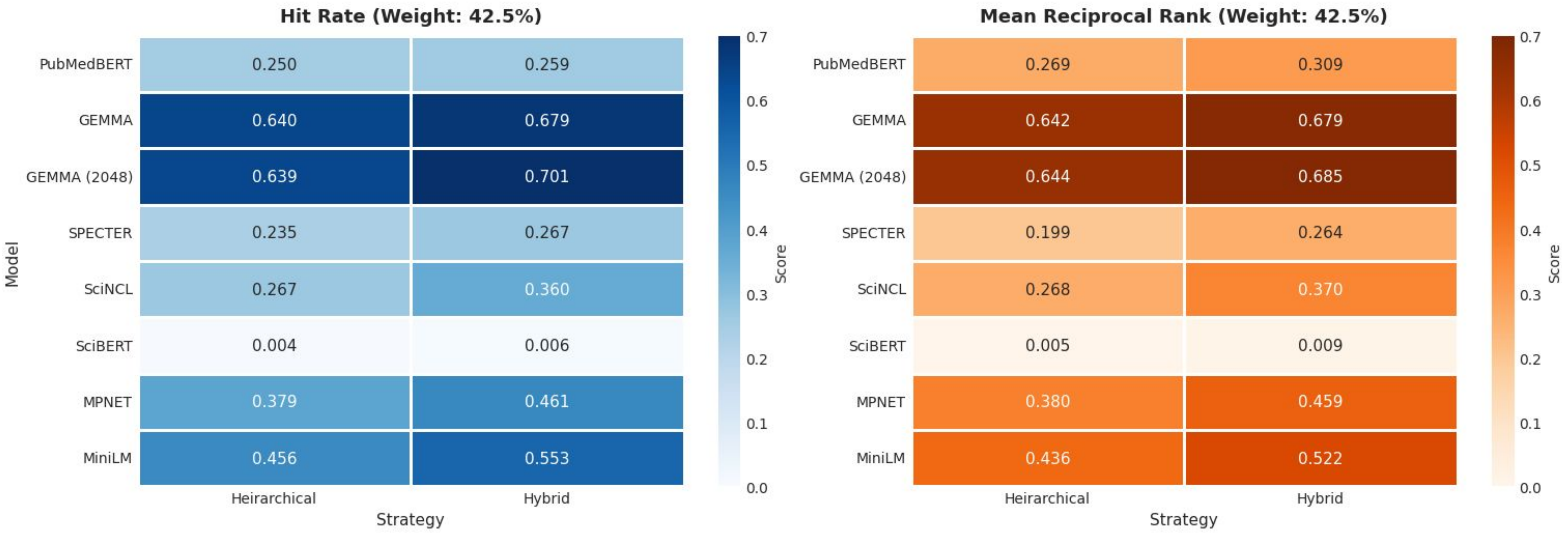
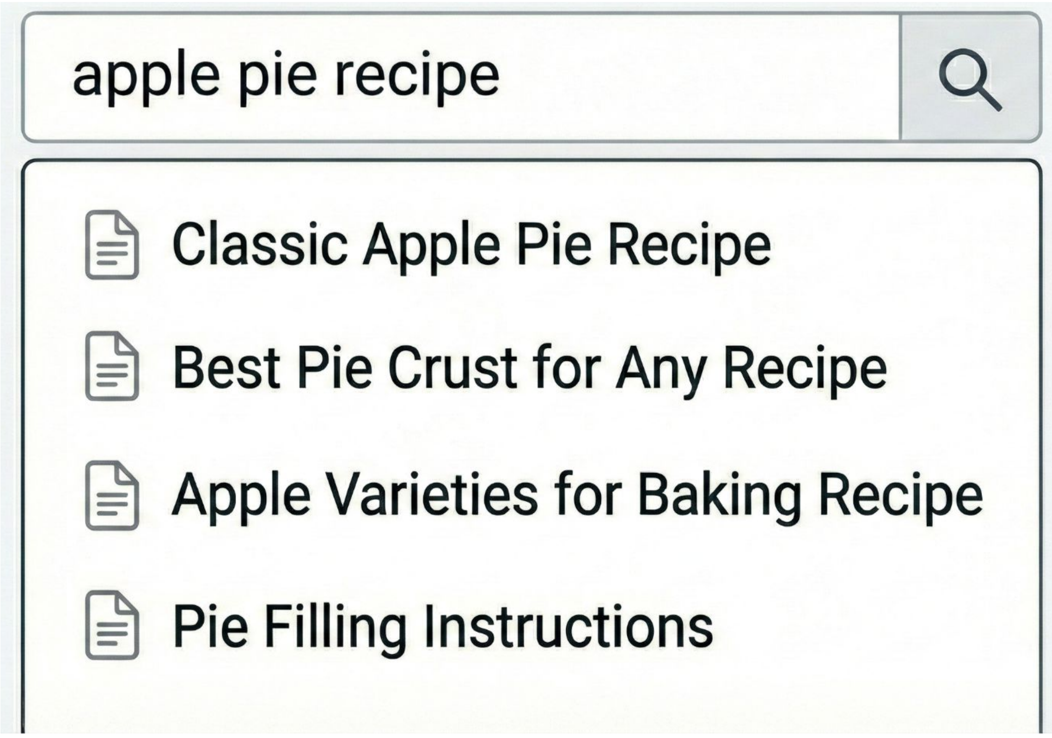


Figure 2: Comparison of the quality metrics across embedding models & chunking strategies

Retrieval & Reranking

Retrieval/ Search

- Method to fetch relevant information
- Keyword - Through sparse vector search
- Semantic - Meaning based through dense vector search

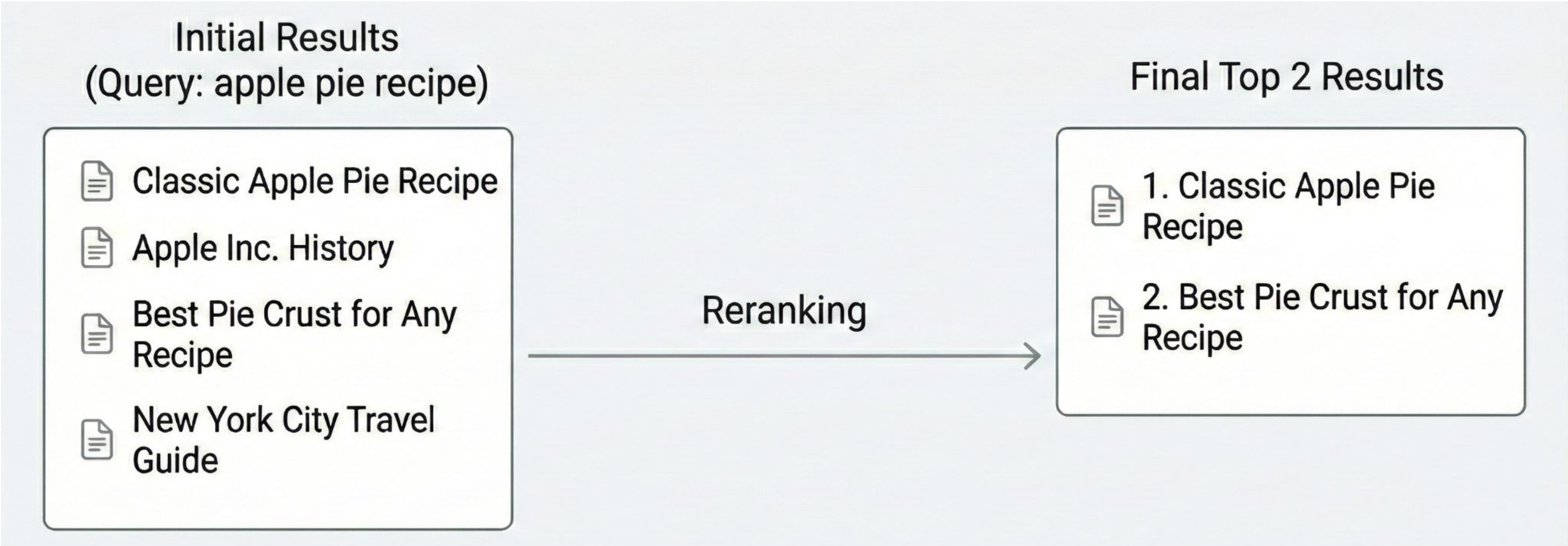


Keyword Search

Semantic Search

Reranking

- Prioritization of retrieved data
- Coarse retrieval to fine reranking



Reranking

MS-MARCO Reranker Is Most Efficient

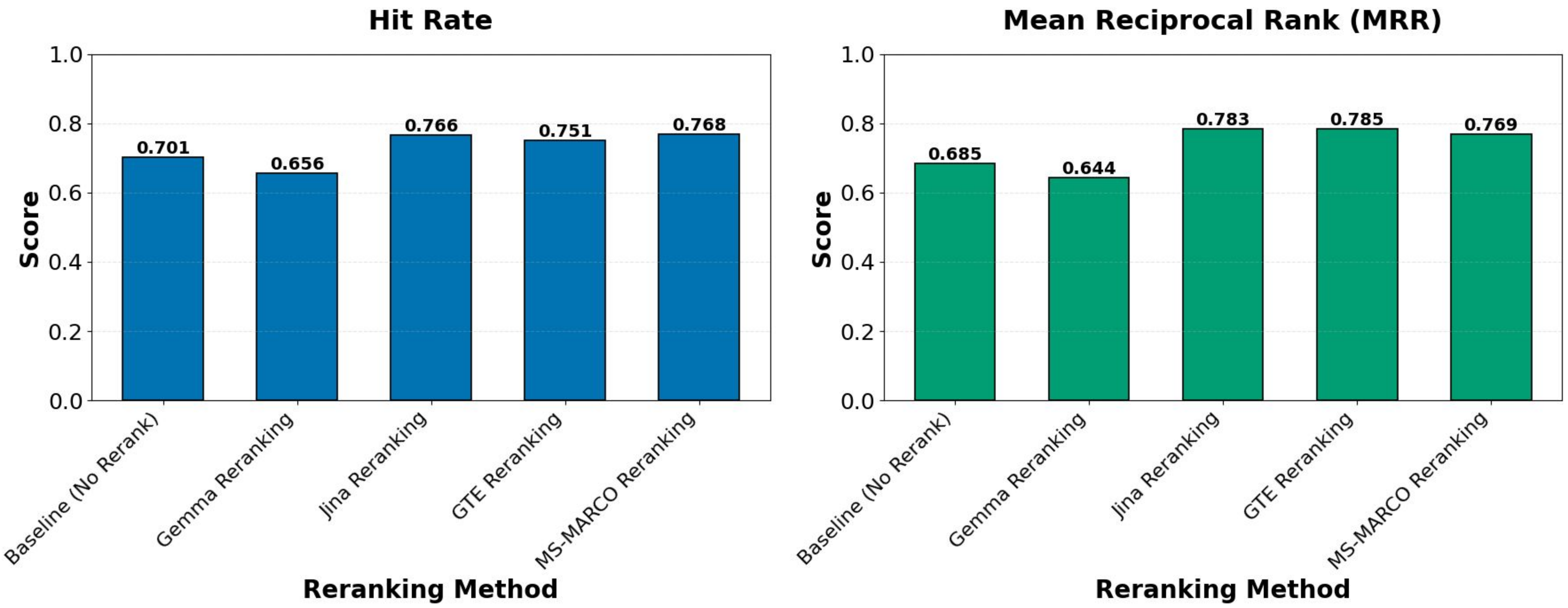


Figure 3: Comparison of the quality metrics of reranking models

Keyword Search with Reranking Has Highest Scores

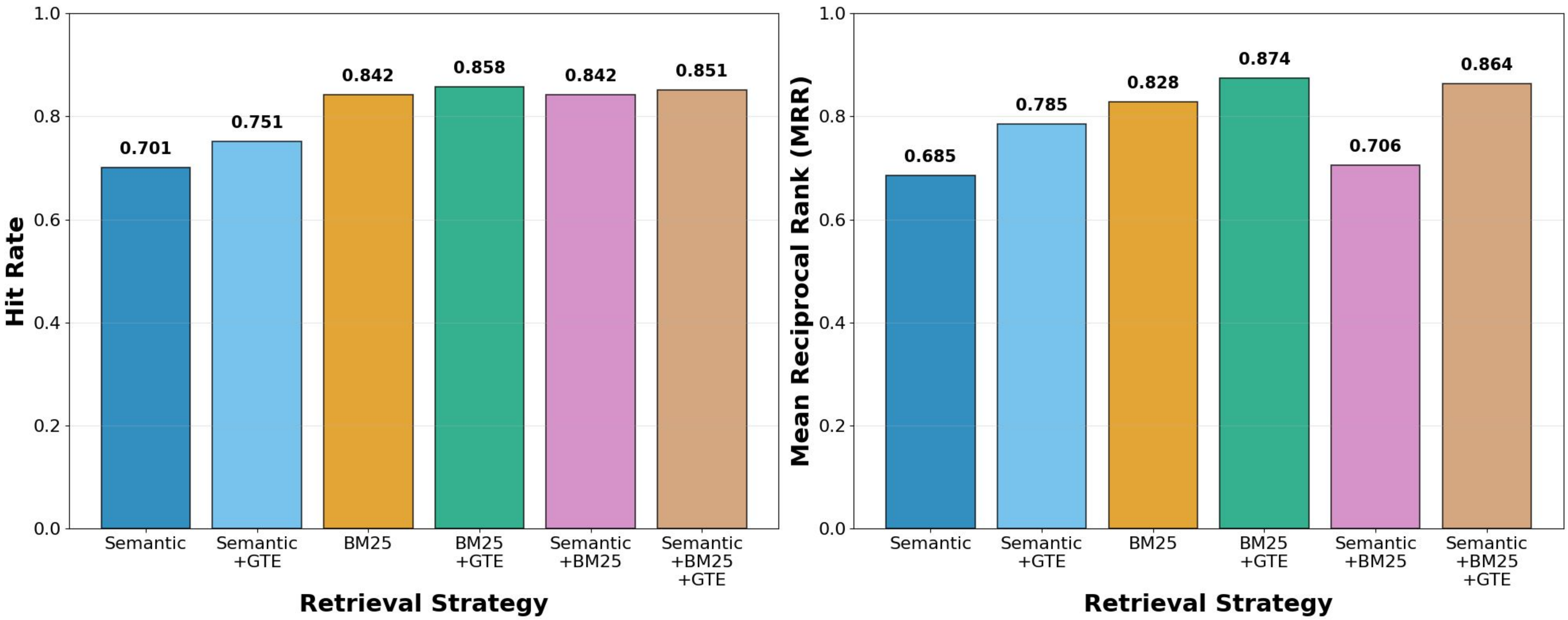


Figure 3: Comparison of the quality metrics across various retrieval strategies

Generation Model

Role

- Formulate answers from retrieved scientific paper segments
- Decide next course of action (e.g., retrieve more information,)

Requirements

- Hardware: ≤20GB VRAM
- Deployment: Local / on-premises
- Functional:
  - Capable of conversation, answer synthesis, moderate reasoning
  - Tool calling

Candidates: Qwen3 8B, Llama3.1 8B, Qwen3 VL 8B

Faithfulness & Relevancy

Query: What is the typical outcome of a MALDI-imaging study?

Retrieved: A typical MALDI-imaging study results in a set of ions of interest

Generated Answer: MALDI-imaging studies produce ion distribution maps. The technique requires extensive sample preparation.

Faithfulness = supported claims / total claims

Claim	Supported?
Produces ion maps	✓
Requires extensive prep	✗

Faithfulness = 0.5

Relevancy = similarity(query, questions generated from answer)

Claim	Similarity
What do MALDI studies produce?	0.82
What does the technique require?	0.34

Relevancy = 0.58

Qwen3 8B Has Good Balance of Faithfulness & Relevancy

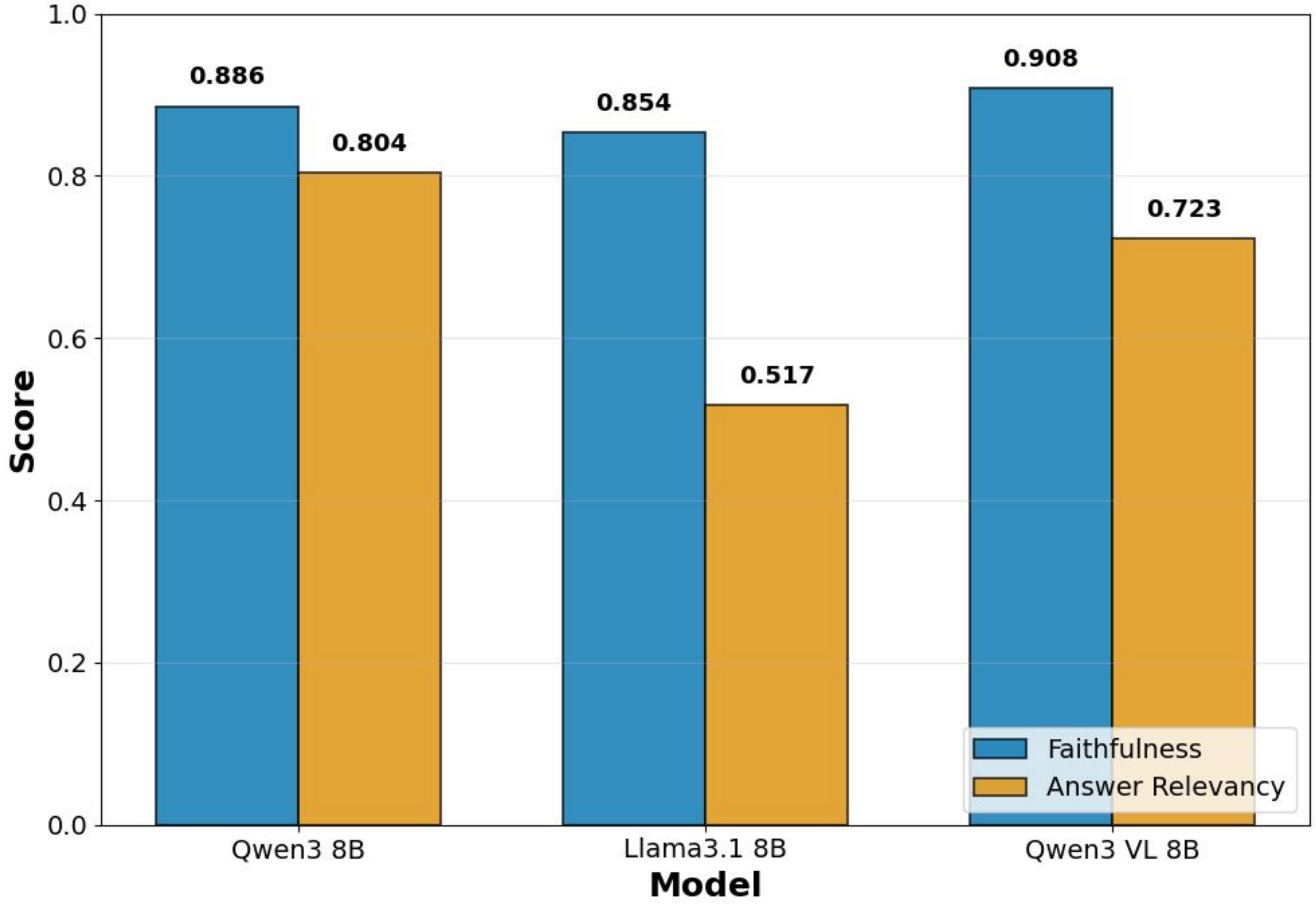


Figure 4: Comparison of generation model quality metrics across on RAG workflow

Hallucination

Retrieval Based Queries

- Output not grounded in retrieved documents

Query: Where did Mr. Bob live?
Documents: Mr. Bob loves to travel. He lives in USA.

Hallucination: Mr. Bob lives in Massachusetts.

Fact Based Queries

- Information the model hasn't seen (low confidence.)

Query: What's Adam Kalai's birthday?

Hallucination: Adam Kalai's birthday is March 7th.

Brainstorming Queries

- No shackles

Query: Can moss predict internet outages?

????: Track their growth rate pattern variations.

Hallucination Detection - MiniCheck RoBERTa Has Largest Efficiency

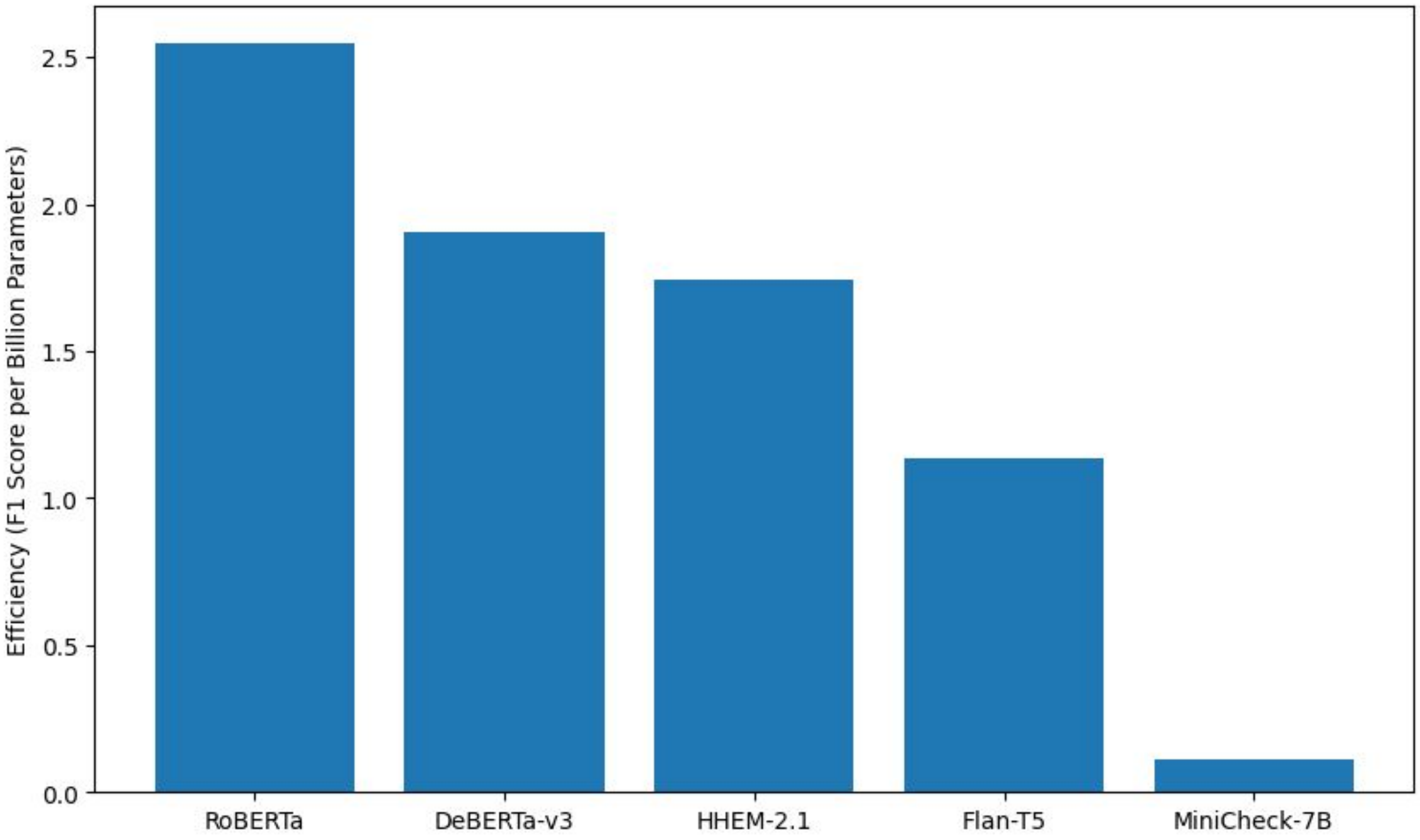


Figure 5: Efficiency comparison of the hallucination detection models

Optimal F-1 Score 85.3% At Threshold 26.3%

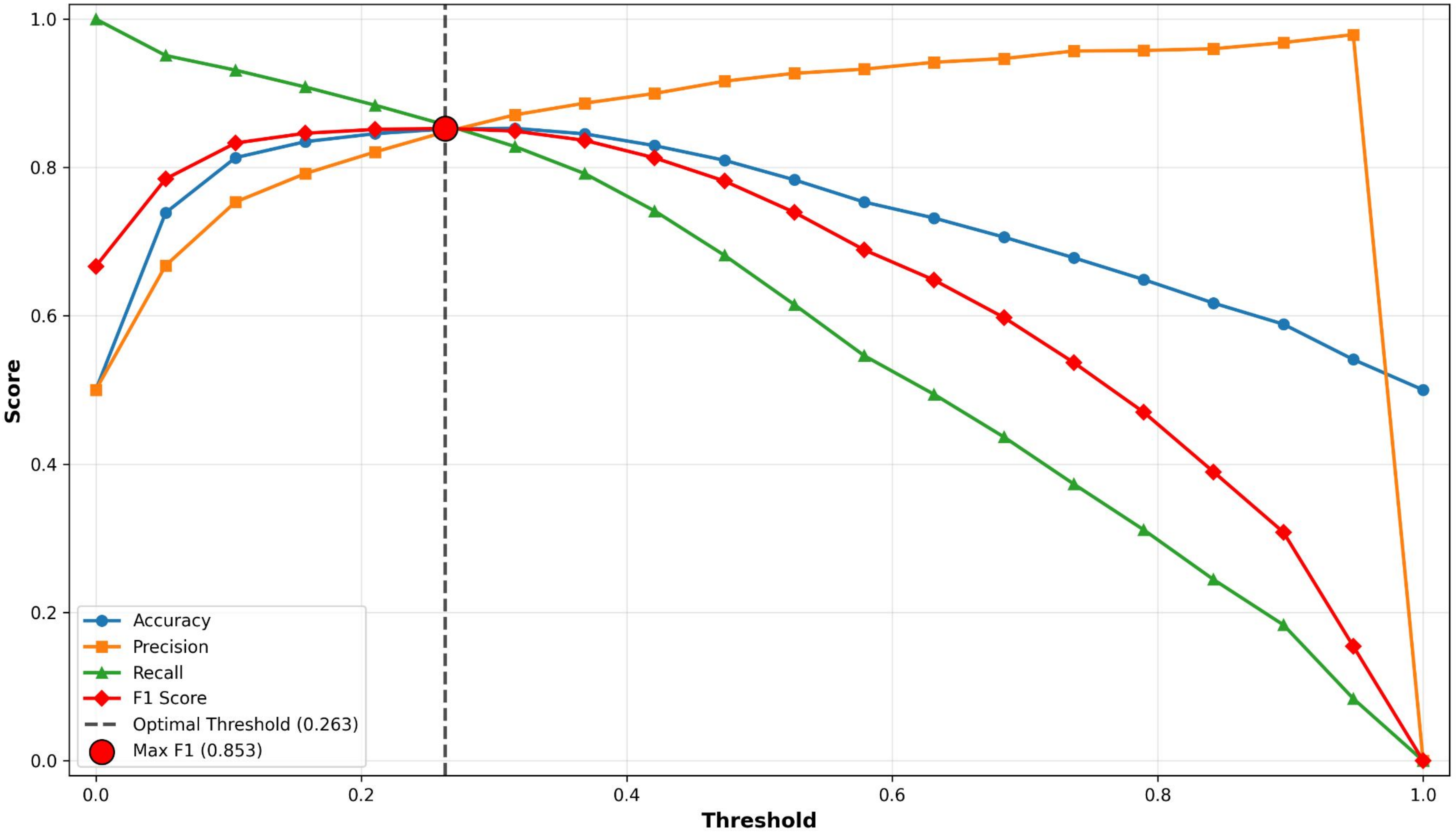


Figure 6: Confidence threshold gridsearch across accuracy, precision, recall, and f1 score on RoBERTa model

Three-Tiered Hallucination Reporting

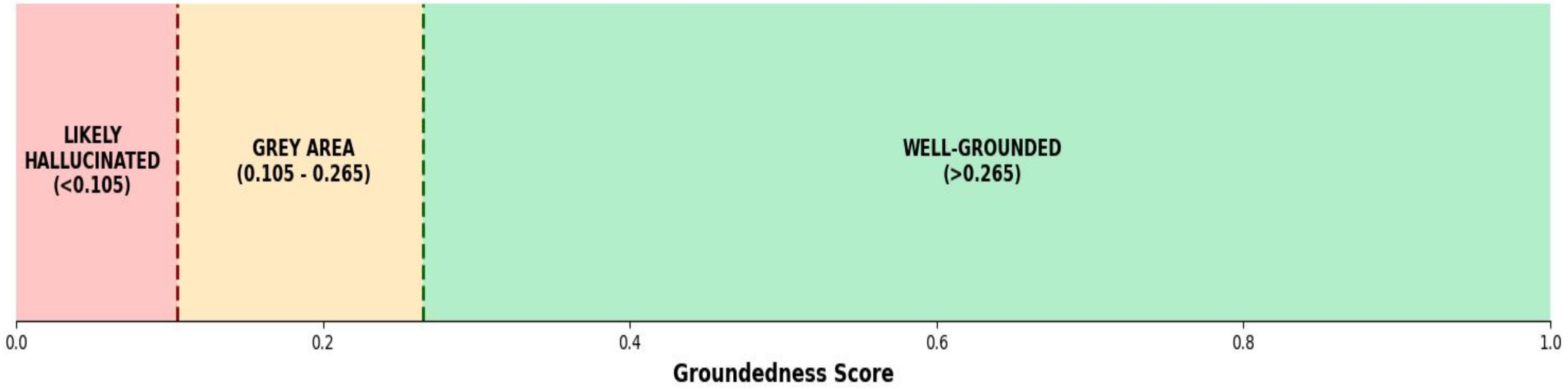


Figure 7: Hallucination detection and reporting decision framework

Hallucination Mitigation Strategy - Confidence Thresholding

Hypothesis: Explicit instruction thresholding reduces hallucination rate

Example: Answer only if you are t = 80% confident. Mistakes cost you t/(1-t) points

Synthetic Data Generation

- Query:
  - Answerable
  - Gray Area
  - Unanswerable
- Excerpt Answer

Inferencing

Inference chosen chat model with threshold and non-threshold prompt

Validation

Validate if the differences are significant

Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why Language Models Hallucinate. arXiv preprint arXiv:2509.04664.

Confidence Thresholding Prompt Types

Prompt Type	**Core Instruction**
Baseline	[No instructions] — Just question + context
Explicit IDK	"If the context doesn't contain the answer, say 'I don't know'"
Confidence Threshold	"Answer only if ≥80% confident. Mistakes cost 4 points, correct = 1 point, IDK = 0"
Confidence Rubric	"Check 5 criteria (answer explicit, info complete, no ambiguity, etc.). Answer only if 4/5 satisfied"

Confidence Rubric Handles Ambiguity Well

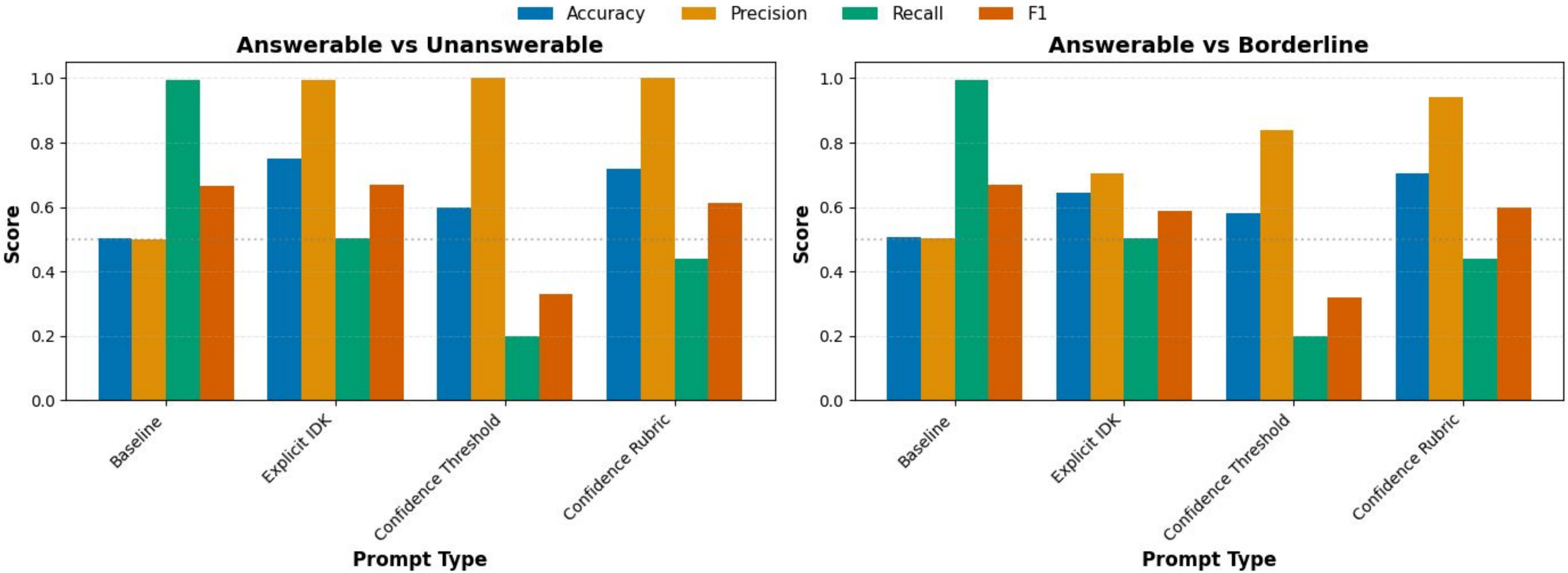


Figure 8: Impact of prompting strategies on hallucination in Qwen3 8B

Confidence Thresholding Key Takeaways

- **Baseline**: Answers everything, including unanswerable queries
- **Explicit IDK:** Best precision-recall tradeoff for clear queries, but precision drops ~29% on ambiguous queries
- **Confidence Rubric:** Most robust to ambiguity (only ~6% precision drop)
- **Confidence Thresholding:** High precision but overly conservative (~20% recall)

Hallucination Mitigation Strategy - Context Management

Premise: Beyond certain context length performance degrades

Objective: Find optimal context window to minimize hallucination

Application: Limit conversation length or apply context management beyond optimal context window

Synthetic Data Generation & Processing

- ➢ Query
- ➢ Excerpt Answer
- ➢ Context Padding:
  - ○ Top
  - ○ Bottom
  - ○ Middle

Inferencing

Inference chosen chat model

Validation

- ➢ Validate if the differences are significant
- ➢ Obtain optimal context window

Chroma Research. (2025). Context Rot: How Increasing Input Tokens Impacts LLM Performance. Retrieved from https://research.trychroma.com/context-rot

Increased Context Length Leads to Fewer Responses

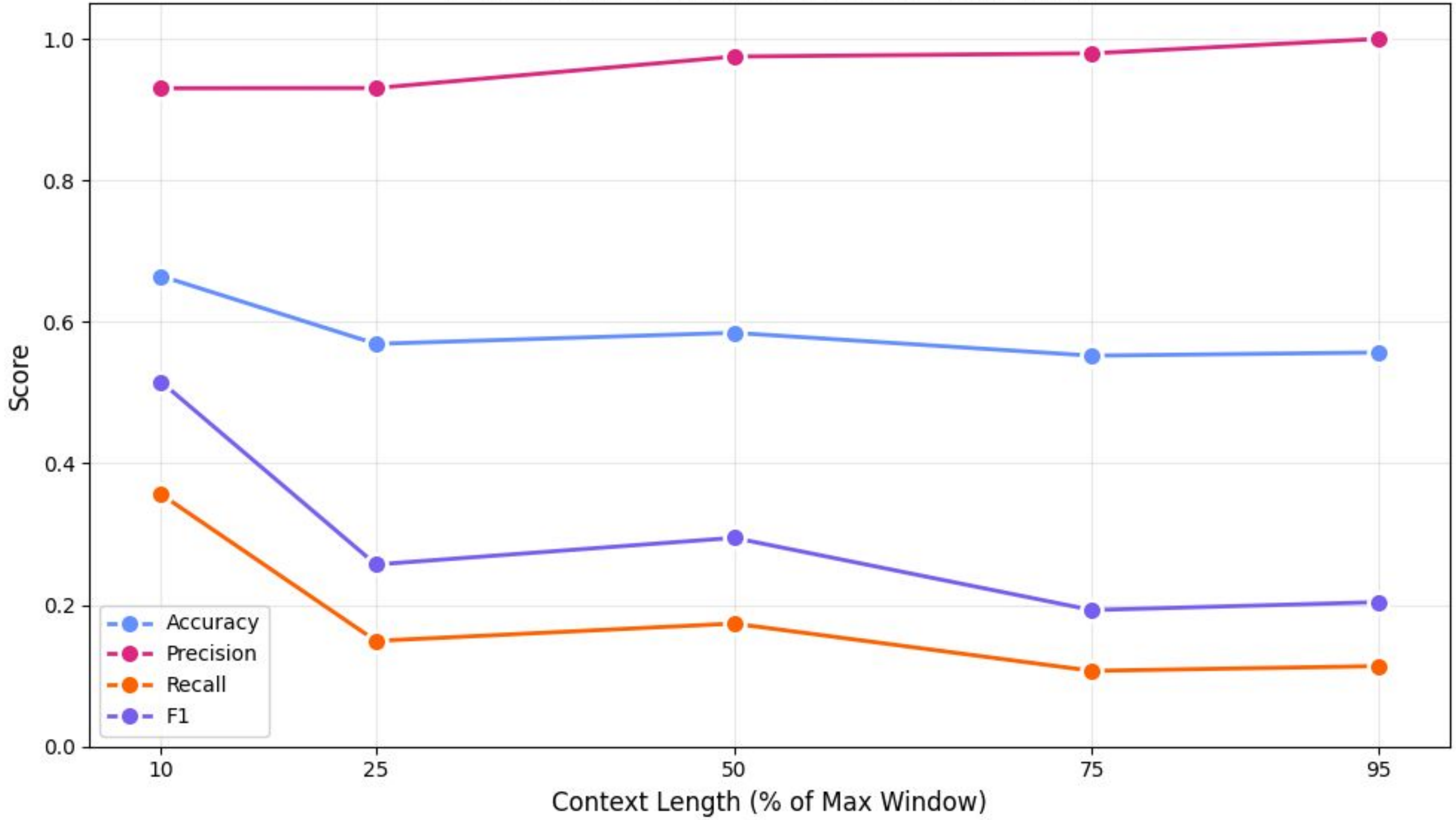


Figure 9: Performance of Qwen3 8B across context lengths on answerable and unanswerable questions

Answer Gets Lost in the Middle

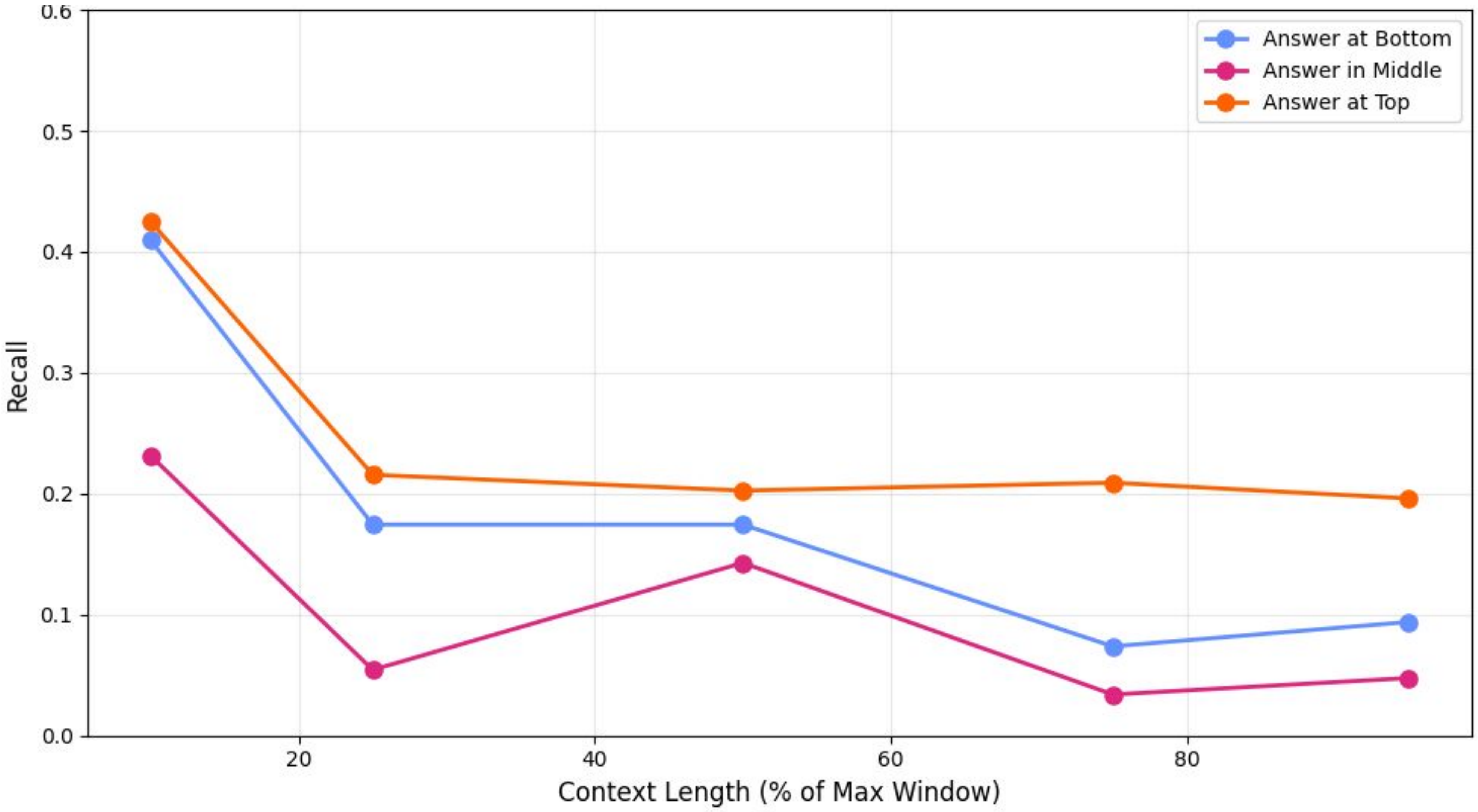


Figure 10: Comparison of the recall across three answer locations: top, bottom, and middle

Context Management Key Takeaways

- Qwen3 8B becomes overly conservative at longer context lengths
- Recall drops - model fails to find answers buried in long contexts
- "Lost in the Middle" phenomenon - important info if present in the middle gets overlooked
- **Recommendation**:
  - Limit context to ~3.2K tokens (10%) or extensive context management (E.g, Compression)
  - Front-load critical reference information

Project Scorecard

Objective	**Component**	**Target**	**Status**	**Result**
Queryable Repository	Parsing, Chunking, Embedding, Retrieval	Hit Rate@10 ≥75% MRR@10 ≥65%	✓	Hit Rate@5 = 85.1% MRR@5 = 86.4%
	Chat Model	Faithfulness ≥85% Relevancy ≥80%	✓	Faithfulness = 88.6% Relevancy = 80.04%
Private	GPU Memory	≤25GB VRAM	✓	~18GB VRAM
	Latency	Simply Query: <10s Complex Query: <60s	⚠️	-
	External API	None	✓	Fully private
	Deployment	Integrate with Slack	⚠️	-
Groundedness	Hallucination Detection	F1 ≥80%	✓	F1 = 85.3%
	Hallucination Mitigation	Precision ≥85%	✓	Precision = 93%

Discussion

- Docling provided quality text extraction and structure retainment
- Higher potential configurations chosen over top quality metrics
- User testing necessary - Synthetic query vs Real query discrepancy
- Hallucination mitigation precision-recall tradeoff challenging
- Moving forward, explore architecture & deployment

Thank You