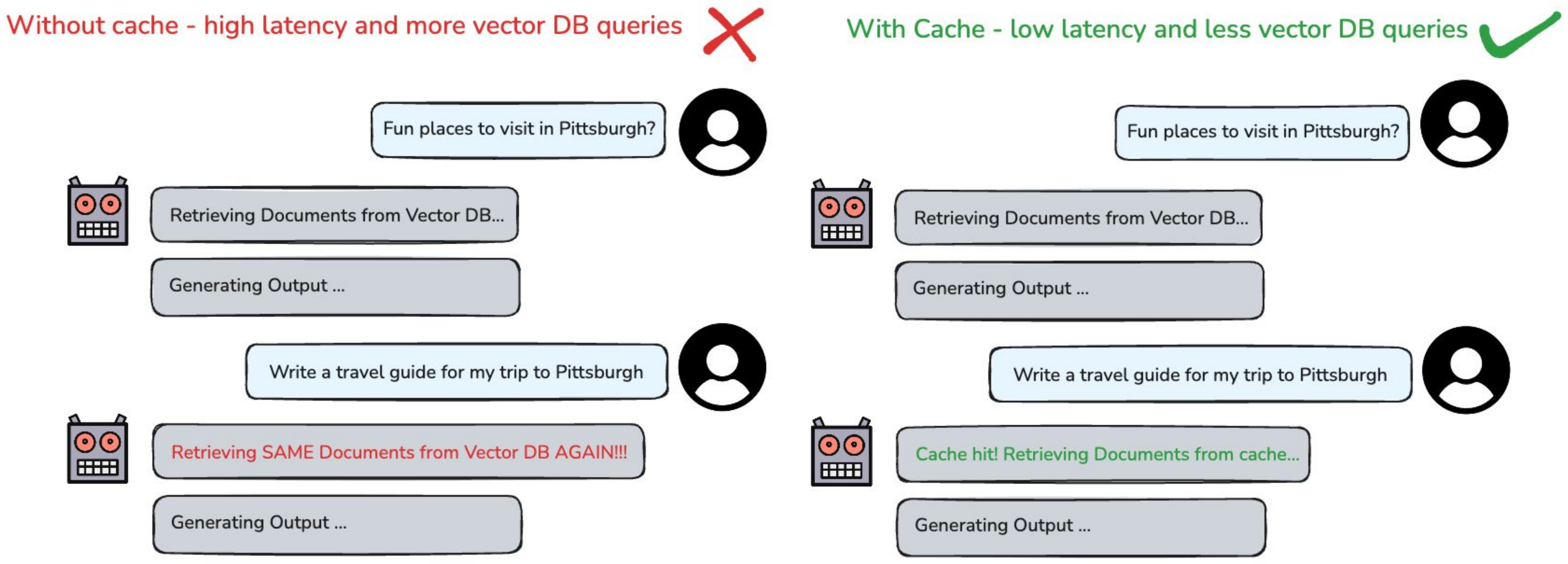
Retrieval Optimization with Semantic Cache

Ziming Wang, Tim Han, Dingfan Zheng, Lei Li
{peterwa2,timhan,dingfanz,leili}@cs.cmu.edu
Carnegie Mellon University

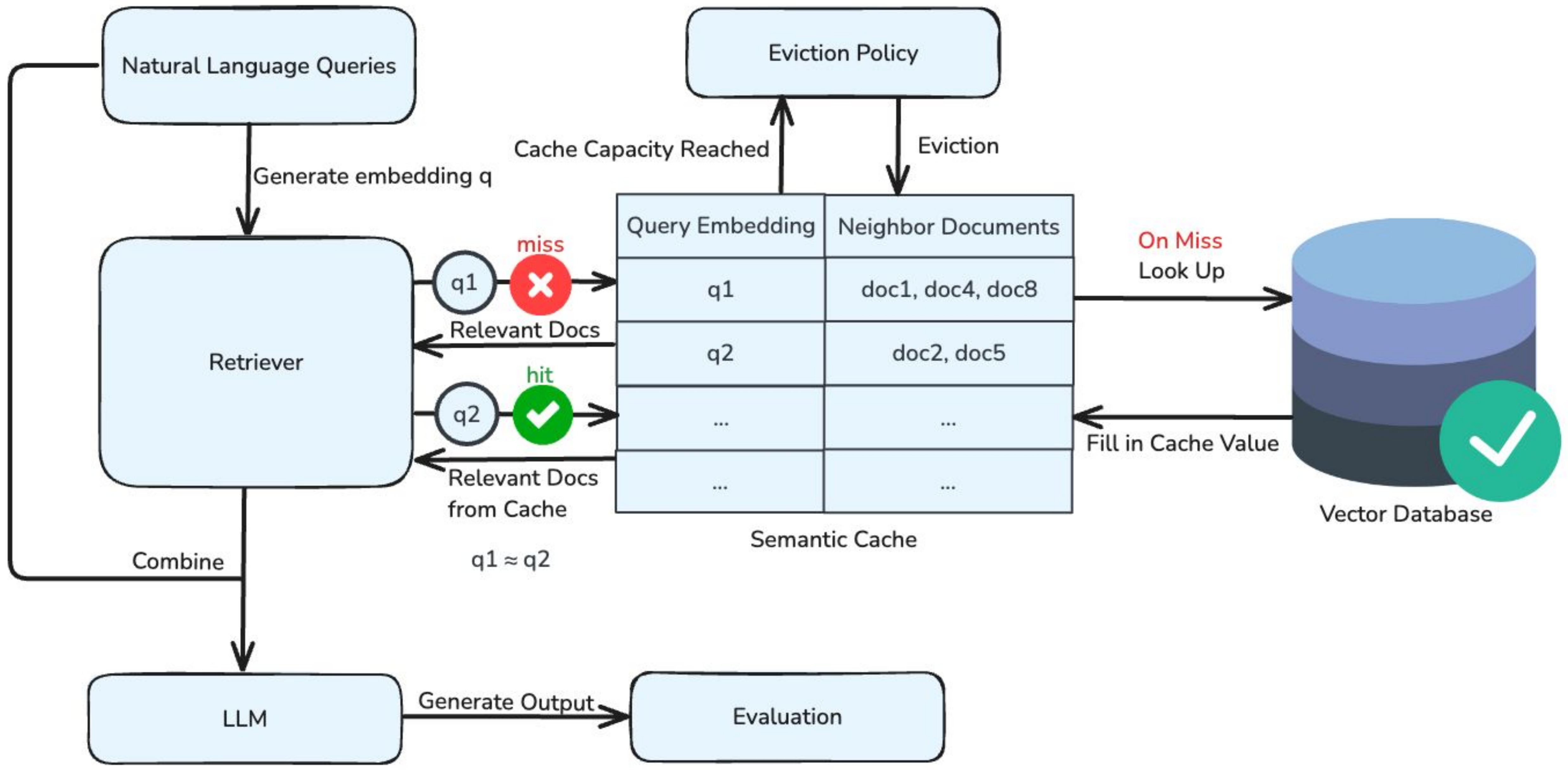
Caching in LLM RAG Applications

- Previous work caches LLM outputs, not fine-grained enough
- Semantically similar question retrieve similar documents, cache the query and retrieved document instead!


Without cache - high latency and more vector DB queries
Fun places to visit in Pittsburgh?
Retrieving Documents from Vector DB...
Generating Output ...
Write a travel guide for my trip to Pittsburgh
Retrieving SAME Documents from Vector DB AGAIN!!!
Generating Output ...
With Cache - low latency and less vector DB queries
Fun places to visit in Pittsburgh?
Retrieving Documents from Vector DB...
Generating Output ...
Write a travel guide for my trip to Pittsburgh
Cache hit! Retrieving Documents from cache...
Generating Output ...

Semantic Cache System Design

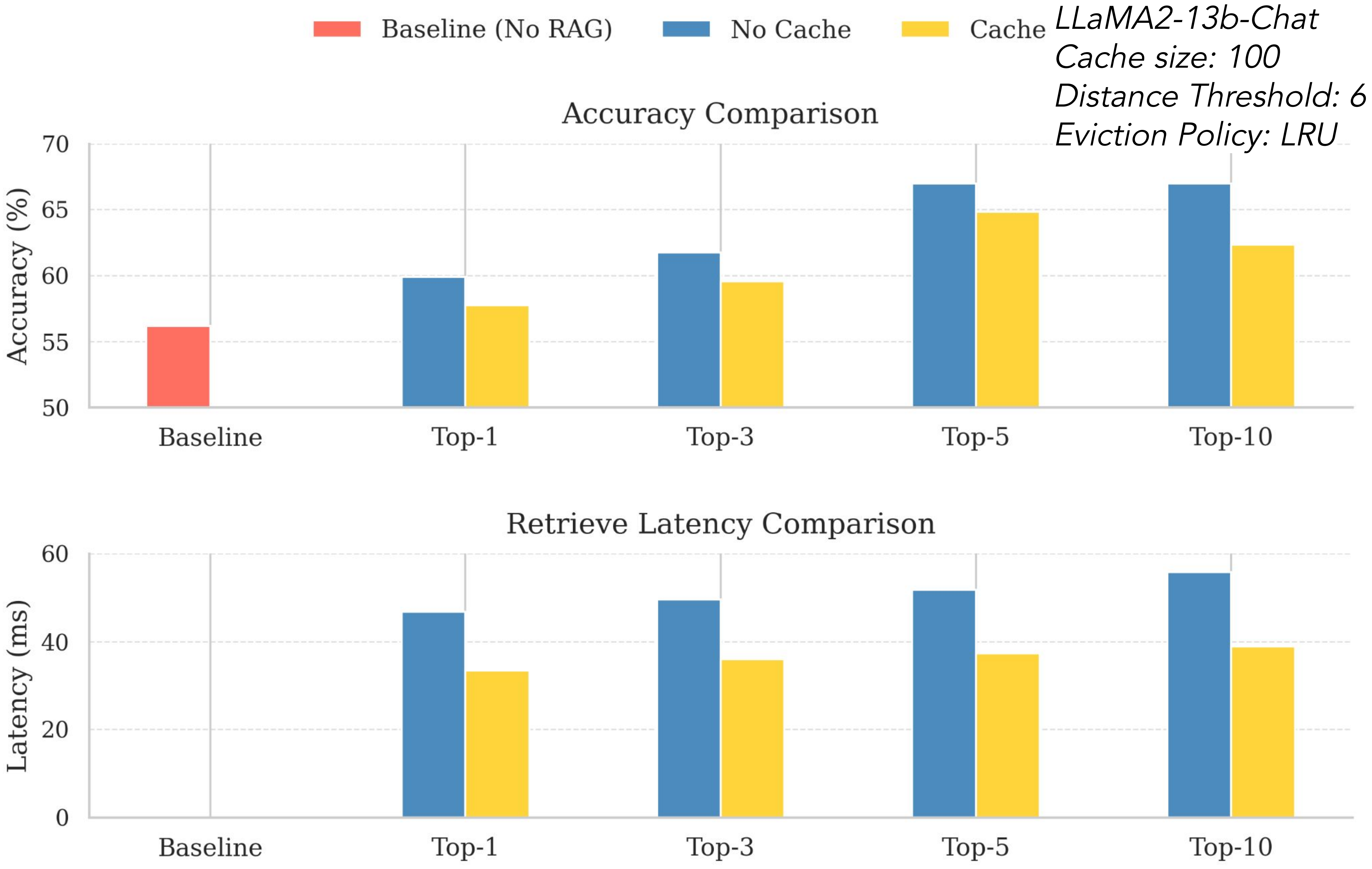
- Reuse topk results for similar queries based on distance threshold.
- Tunable cache size, threshold, and eviction policies for optimization.



Query Embedding	Neighbor Documents
q1	doc1, doc4, doc8
q2	doc2, doc5
...	...
...	...

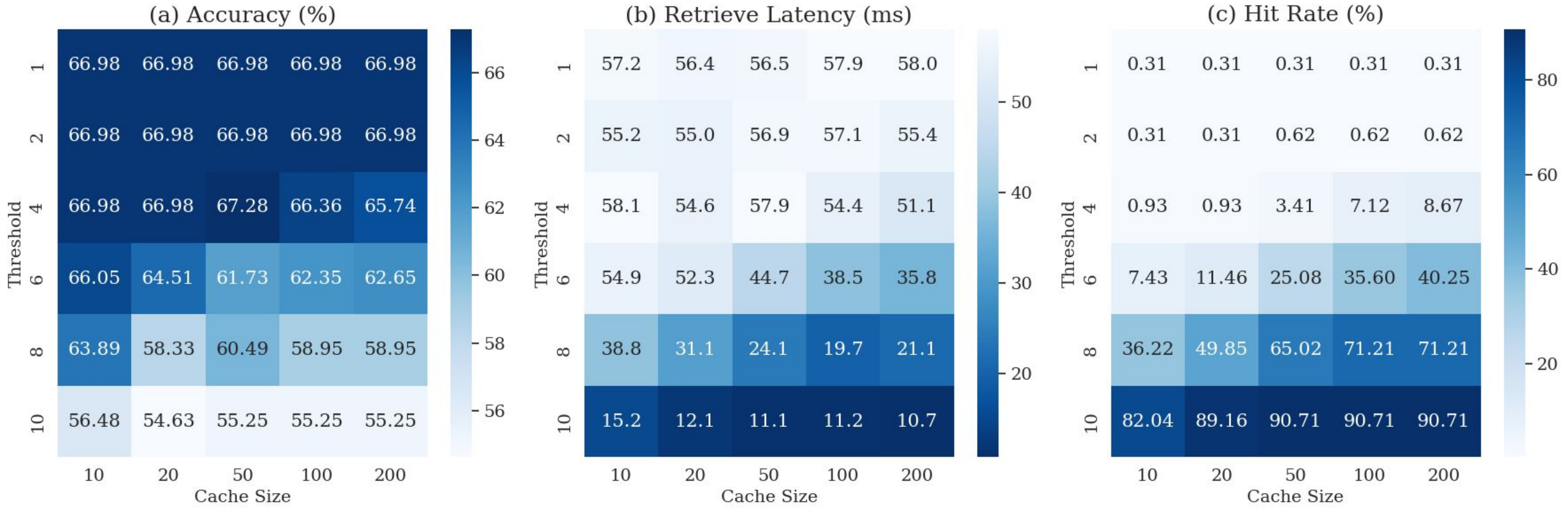
Results

- Our approach cuts retrieval latency by > 30%, while maintaining high retrieval quality with only a 2–4% drop in accuracy.



LLaMA2-13b-Chat
Cache size: 100
Distance Threshold: 6
Eviction Policy: LRU

- Impact of Cache Size and Threshold on Performance Metrics



Key Takeaway / Advantages

- Reduces retrieval latency and the number of API calls
- Tuning cache configs based on workloads and resources