Vector Database Indexing Benchmark

Weaviate HNSW vs Exact Baseline

This report presents a comparative analysis of approximate
and exact vector indexing strategies under realistic dataset scale.

Dataset Description

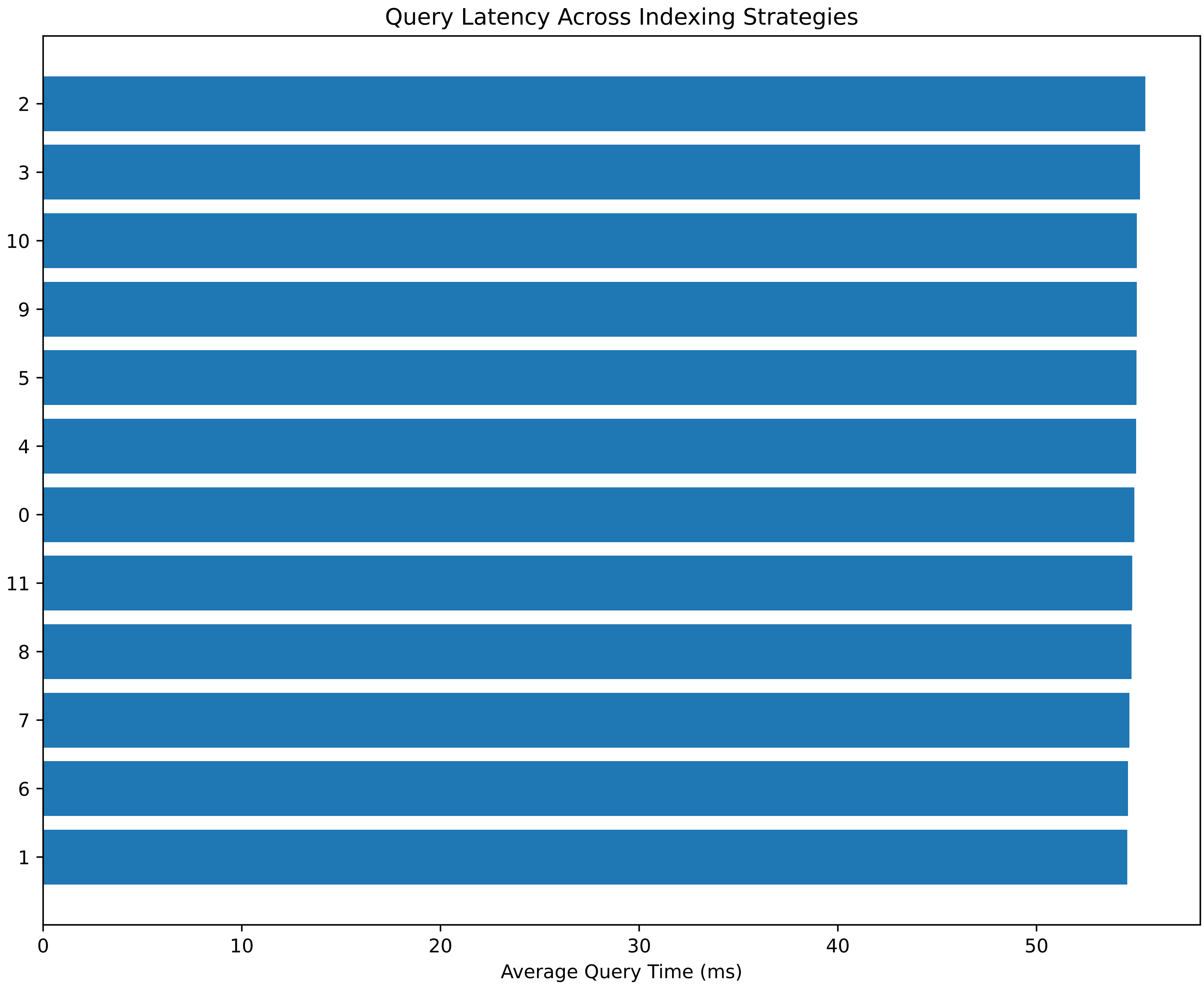
- Dataset: MuskumPillerum / General-Knowledge
- Corpus size: ~10,000 documents
- Query count: ~1,000 questions
- Domain: Open-domain general knowledge

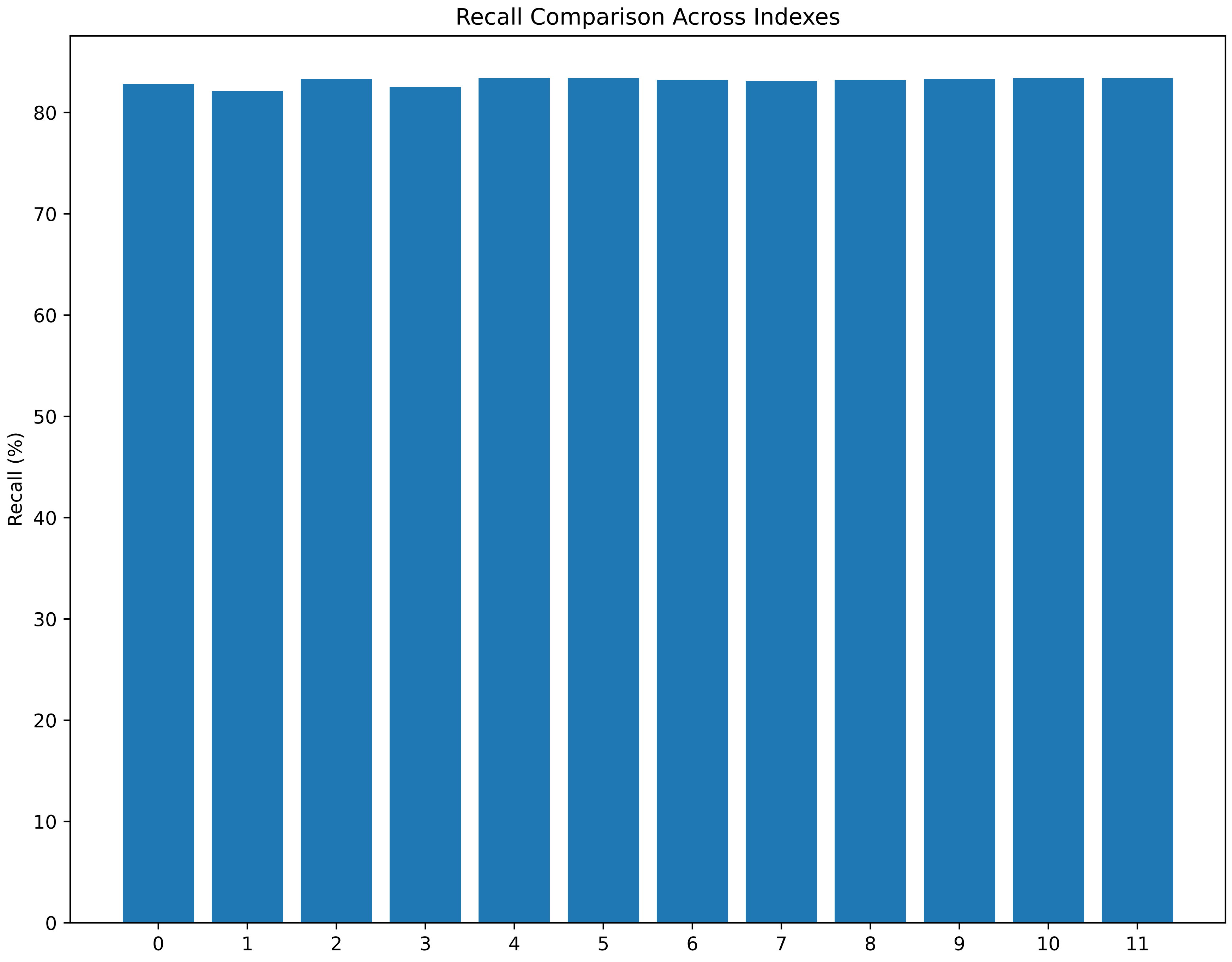
Empty and invalid entries were filtered during preprocessing to ensure robust embedding generation.

1. Exact Baseline Index
   - Brute-force cosine similarity
   - Provides upper bound on recall
   - Computationally expensive

Indexing Strategies Evaluated

2. HNSW (Weaviate)
   - Approximate nearest neighbour graph
   - Tunable parameters: ef, efConstruction, maxConnections
   - Optimised for low-latency retrieval


Query Latency Across Indexing Strategies
2
3
10
9
5
4
0
11
8
7
6
1
0
10
20
30
40
50
Average Query Time (ms)


Recall Comparison Across Indexes
Recall (%)
0
10
20
30
40
50
60
70
80
0
1
2
3
4
5
6
7
8
9
10
11

Observations and Trade-offs

- Exact search achieves maximum recall at significantly higher latency.
- HNSW configurations offer substantial latency reductions with marginal recall degradation.
- Parameter tuning enables balanced trade-offs depending on workload.

These results demonstrate the necessity of approximate indexing strategies for scalable LLM-backed retrieval systems.