The Recall Fallacy : Benchmarking the Downstream Impact of ANN Approximation on Large-Scale Dense Retrieval

Samip A. Singhal
Department of Computer Science, New York University
Information Retrieval, Fall 2025

Abstract

Approximate Nearest Neighbor (ANN) search is a foundational component of modern dense retrieval and retrieval-augmented generation (RAG) systems. ANN algorithms are commonly evaluated using geometric recall metrics that measure how closely approximate results match exact nearest neighbors in embedding space. This evaluation paradigm implicitly assumes that higher ANN recall directly translates to improved retrieval quality.

In this work, we challenge this assumption and introduce the Recall Fallacy: the belief that maximizing ANN recall is always the correct optimization objective for dense retrieval systems. Using the MS MARCO Passage Ranking dataset with 8.8 million passages and human relevance judgments, we benchmark five ANN indexing strategies spanning exact search, inverted-file indices, graph-based indices, and hybrid designs. We analyze the relationship between geometric recall, downstream ranking effectiveness (NDCG@10, MRR@10, MAP@100), and system cost (latency, throughput, and memory footprint). Our results show that ranking quality saturates well before perfect ANN recall is achieved, while computational cost continues to increase sharply. Graph-based ANN methods dominate the speed–accuracy trade-off, while aggressive compression introduces an irreducible quality ceiling.

1. Introduction

Dense retrieval systems encode queries and documents into a shared embedding space and retrieve candidate documents using Approximate Nearest Neighbor (ANN) search. This architecture underpins modern web search, question answering, and retrieval-augmented generation (RAG) pipelines. At large corpus scales, exact nearest neighbor search is computationally infeasible, making ANN algorithms a practical necessity.

Despite their central role, ANN methods are typically evaluated using geometric recall metrics such as Recall@K, which quantify how many exact nearest neighbors are recovered by an approximate index. This practice implicitly assumes that geometric proximity in embedding space aligns with human relevance. In reality, many geometrically close neighbors are irrelevant, while relevant documents may not be among the nearest vectors.

This mismatch motivates a fundamental question: does improving ANN recall meaningfully improve retrieval quality as measured by ranking metrics? In this work, we empirically study the relationship between ANN approximation quality and downstream retrieval effectiveness on a large-scale dataset with human relevance judgments.

2. Scope and Design Choices

This project focuses explicitly on large-scale document retrieval, not general-purpose ANN benchmarking. All experiments are conducted under a dense retrieval paradigm using Maximum Inner Product Search (MIPS), implemented as cosine similarity over L2-normalized embeddings. This reflects standard practice in neural information retrieval.

To isolate the impact of ANN approximation, we fix the embedding model across all experiments and vary only ANN index structures and search parameters. Evaluation emphasizes downstream ranking metrics and system-level trade-offs, rather than geometric recall alone.

3. Experimental Setup

3.1 Dataset

We use the MS MARCO Passage Ranking dataset, consisting of approximately 8.8 million passages and thousands of natural language queries with human relevance judgments. Evaluation is performed on the MS MARCO development set (6,980 queries).

3.2 Embeddings

Queries and documents are encoded using the sentence-transformers/all-mpnet-base-v2 bi-encoder, producing 768-dimensional dense embeddings. All embeddings are stored in float32 format and L2-normalized, enabling cosine similarity to be computed as inner product.

4. Cloud Infrastructure and Engineering Complexity

Due to the scale of the dataset, all experiments were conducted on Google Cloud Platform (GCP) using a bifurcated infrastructure strategy: GPU-accelerated instances for embedding generation and high-memory CPU instances for ANN indexing and evaluation.

4.1 Embedding Generation (GPU)

- **Machine Type**: g2-standard-8
- **Accelerator**: NVIDIA L4 GPU (24 GB)
- **Compute**: 8 vCPUs, 32 GB RAM

The corpus was partitioned into 44 shards of approximately 200k passages. Embeddings were generated using multi-process inference, achieving an average throughput of approximately 2,500 passages per second. Shards were written incrementally to disk to ensure fault tolerance and enable resume-on-failure.

4.2 Indexing and Evaluation (CPU)

- **Machine Type**: n2-highmem-32
- **Compute**: 32 vCPUs
- **Memory**: 256 GB RAM

High memory capacity was required to store raw embeddings (~26.5 GB), graph structures for HNSW-based indices, and per-query search buffers. FAISS was compiled with OpenMP and configured to use 32 threads for all index construction and evaluation runs.

5. ANN Methods Evaluated

We evaluate five ANN configurations implemented using FAISS, spanning the major design paradigms used in practice:

- **FlatIP (Exact Oracle)**
  Brute-force inner product search. Serves as the ground truth for geometric recall and an upper bound on retrieval quality.
- **IVF-Flat**
  Inverted file index with exact vector storage. Search exhaustiveness is controlled via nprobe.
- **IVF-PQ**
  Inverted file index with Product Quantization, compressing vectors by over 95% to reduce memory footprint.
- **HNSW (Hierarchical Navigable Small World)**
  Graph-based ANN index enabling efficient navigation via a multi-layer proximity graph. Search depth is controlled via efSearch.
- **IVF-HNSW (Hybrid)**
  A two-stage hybrid index combining coarse inverted-file partitioning with HNSW-based graph search inside selected clusters. This design explores whether coarse pruning and graph navigation can be combined to improve efficiency.

6. Evaluation Methodology

For each ANN configuration, we retrieve the top-200 candidates per query. We report:

- **Geometric Recall@200** relative to FlatIP
- **Ranking Metrics**: NDCG@10, MRR@10, MAP@100
- **Latency**: per-query p95 latency

- **Throughput**: Queries per second (QPS)

Oracle nearest neighbors are computed once using FlatIP and cached for reuse across all experiments.

7. Results

Table 1: Peak Retrieval Performance Across ANN Methods

Method	Configuration	Recall @200	NDCG@ 10	MRR @10	MAP @100	QPS	p95 Latency
FlatIP	Exact	1.000	0.390	0.331	0.338	210	4.90 ms
IVF-Flat	nprobe = 128	0.927	0.382	0.325	0.332	564	1.83 ms
IVF-PQ	nprobe = 128	0.627	0.328	0.275	0.283	2,375	0.43 ms
HNSW	efSearch = 256	0.949	0.386	0.327	0.334	5,948	0.19 ms
IVF-HNSW	nprobe = 32, efSearch = 128	0.842	0.362	0.309	0.315	1,980	0.54 ms

7.1 Observing the Recall Fallacy

Across all ANN methods, ranking quality improves rapidly at low recall levels and then plateaus. For IVF-Flat, increasing Recall@200 from ~0.89 to ~0.93 nearly doubles p95 latency while yielding only marginal improvements in NDCG. Similar saturation behavior is observed for HNSW and IVF-HNSW.

These results empirically demonstrate the Recall Fallacy: recovering the final few percent of nearest neighbors is computationally expensive but contributes little to downstream ranking effectiveness.

7.2 Compression vs. Quality

IVF-PQ achieves the smallest memory footprint and highest throughput, but introduces a **hard ceiling on ranking quality**. Even with aggressive search parameters, IVF-PQ fails to match the ranking effectiveness of uncompressed indices, indicating irreversible information loss from vector quantization.

7.3 Graph-Based and Hybrid Methods

HNSW dominates the speed–accuracy Pareto frontier, achieving both higher throughput and lower latency than inverted-file methods at comparable ranking quality. IVF-HNSW improves over IVF-Flat by incorporating graph search, but does not surpass pure HNSW, suggesting that global graph navigation remains more effective than hybrid pruning at this scale.

8. Index Size and Memory Footprint

Index Type	Configuration	On-Disk Size
FlatIP	Exact	~26 GB
IVF-Flat	nlist = 16k	~26 GB
IVF-PQ	M = 64, 8-bit	~656 MB
HNSW	M = 16	~27 GB
IVF-HNSW	nlist = 16k	~26 GB

Uncompressed indices are dominated by raw vector storage, while IVF-PQ reduces memory usage by over 95% at the cost of ranking quality.

9. Time Budget

Phase	Duration	Hardware
Data Preprocessing	~30 minutes	CPU
Neural Embedding	~5.4 hours	NVIDIA L4 GPU
ANN Index Build	~2.5 hours	High-Mem CPU
Evaluation Run	~15 minutes	CPU

10. Conclusion

This work provides empirical evidence that perfect ANN recall is neither necessary nor sufficient for high-quality dense retrieval. On a large-scale benchmark with human relevance judgments, ranking metrics saturate at moderate recall levels, exposing the Recall Fallacy underlying common ANN evaluation practices. Graph-based ANN methods (HNSW) offer the most favorable balance between retrieval quality and system efficiency, while aggressive compression imposes an irreducible quality ceiling.

Future work should explore disk-based graph ANN methods and the interaction between ANN approximation and downstream reranking or generative models.

11. Git Link

https://github.com/samipsinghal/nearly