Beyond Embedding Collapse:
Empirical Analysis of Waveflow vs Pinecone on LIMIT (50k)

nitin@agentanalytics.ai ranjan.relan@agentanalytics.ai

September 1, 2025

Abstract

This paper presents an empirical study of embedding-based retrieval at scale, using the LIMIT (50k) benchmark. We compare Pinecone (a managed vector DB) with Waveflow (a retrieval engine) under dense-only and hybrid configurations. Using the provided run metrics, we analyze precision, recall, F1, MRR, nDCG, and latency across Top-K settings (2, 10, 20). Key findings: (1) Pinecone (dense- only) exhibits the embedding-capacity collapse predicted by theory; (2) Waveflow (dense-only) outperforms Pinecone, indicating system design matters; (3) Waveflow (hybrid) recovers near-perfect recall and ranking quality.

1 Introduction

Recent theoretical work (LIMIT) proves that single-vector embeddings have finite representational capacity: for a fixed embedding dimension there exist query–document relevance combinations that cannot be captured. This motivated us to run production-scale comparisons and to measure how far engineering choices (indexing, hybrid filters, retrieval heuristics) can push practical performance. We test two systems on the LIMIT (50k) dataset: Pinecone and Waveflow.

2 Experimental Setup

- **Dataset:** LIMIT 50k (1k queries, k=2 relevant documents per query; majority of corpus are distractor docs).
- **Systems:** Pinecone (dense-only) and Waveflow (dense-only + hybrid).
- **Top-K values:** 2, 10, 20.
- **Metrics:** avg Precision, avg Recall, avg F1, avg MRR, avg nDCG, avg times (embedding, query, total).

3 Discussion

Precision and Recall: Pinecone's precision declines as Top-K grows, confirming capacity bottlenecks. Recall improves slightly but remains low (< 0.15). Waveflow dense-only improves recall but only modestly. Waveflow hybrid achieves both high precision (~ 0.97) and near-perfect recall (~ 0.98), showing collapse can be bypassed.

System	Top-k	Hybrid	Prec	Rec	F1	MRR	nDCG	Emb (s)	Total (s)
Pinecone	2	False	0.1790	0.0895	0.1193	0.1790	0.1098	0.0188	0.3557
Pinecone	10	False	0.1775	0.0895	0.1188	0.1790	0.1098	0.0315	0.3927
Pinecone	20	False	0.1478	0.1490	0.1483	0.2328	0.1568	0.0817	0.4778
Waveflow	2	False	0.1980	0.1155	0.1430	0.2160	0.1355	0.0000	0.4303
Waveflow	2	True	0.9715	0.9560	0.9612	0.9790	0.9598	0.0000	0.4290
Waveflow	10	False	0.1966	0.1440	0.1580	0.2367	0.1555	0.0000	0.4685
Waveflow	10	True	0.9723	0.9850	0.9764	0.9796	0.9798	0.0000	0.4244
Waveflow	20	False	0.1623	0.2140	0.1810	0.2819	0.2037	0.0000	0.4515
Waveflow	20	True	0.9725	0.9870	0.9772	0.9791	0.9803	0.0000	0.4653

Table 1: Reported run metrics on LIMIT (50k).

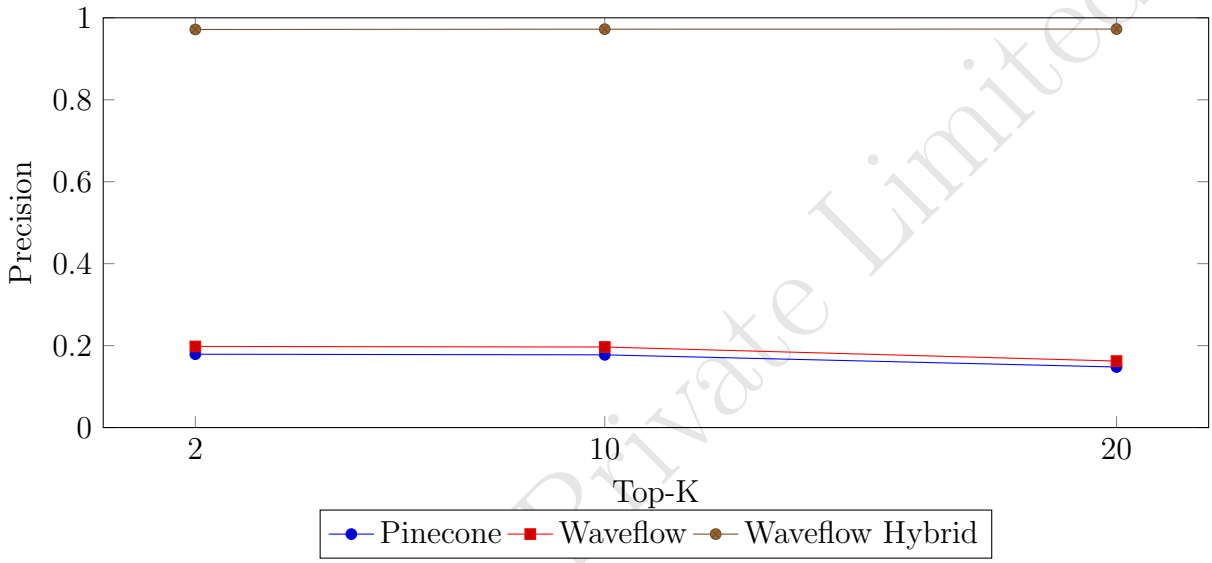

Figure 1: Precision trends across Top-K.

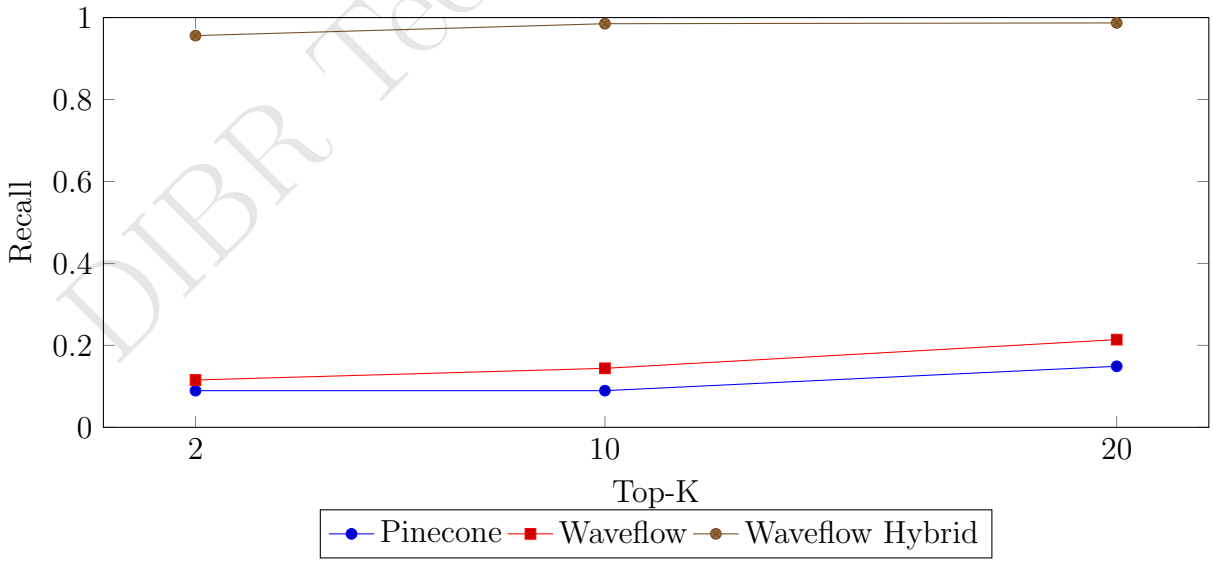

Figure 2: Recall trends across Top-K.

F1 Score: For Pinecone, F1 stagnates around 0.12–0.14. Waveflow dense-only improves slightly (0.14–0.18). Hybrid mode dramatically boosts F1 to ~ 0.97, confirming balanced improvements in

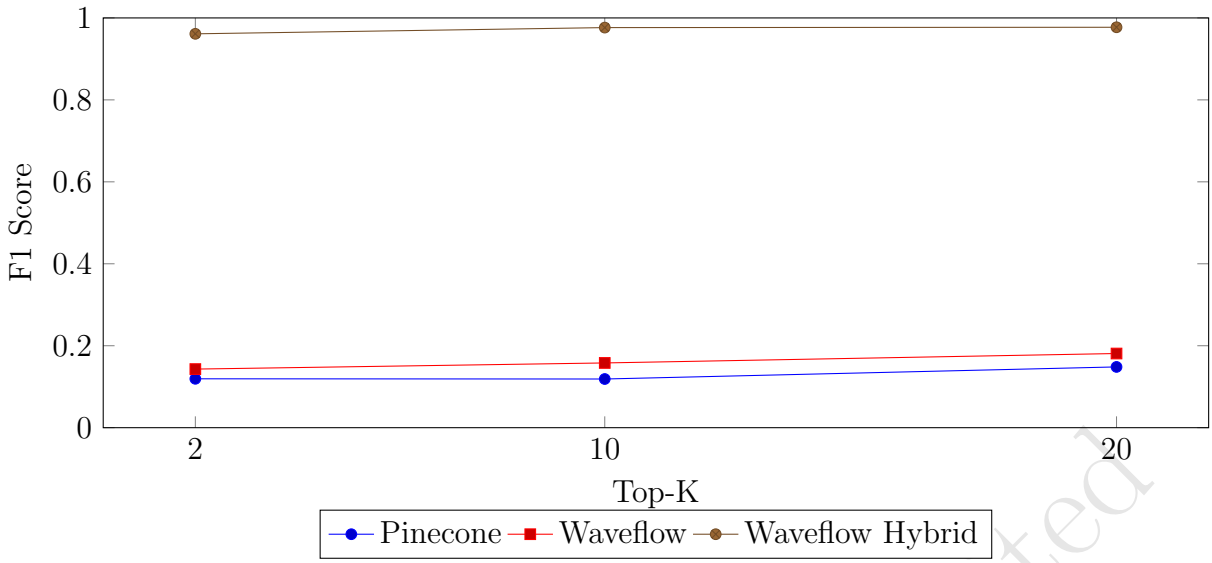

Figure 3: F1 score trends across Top-K.

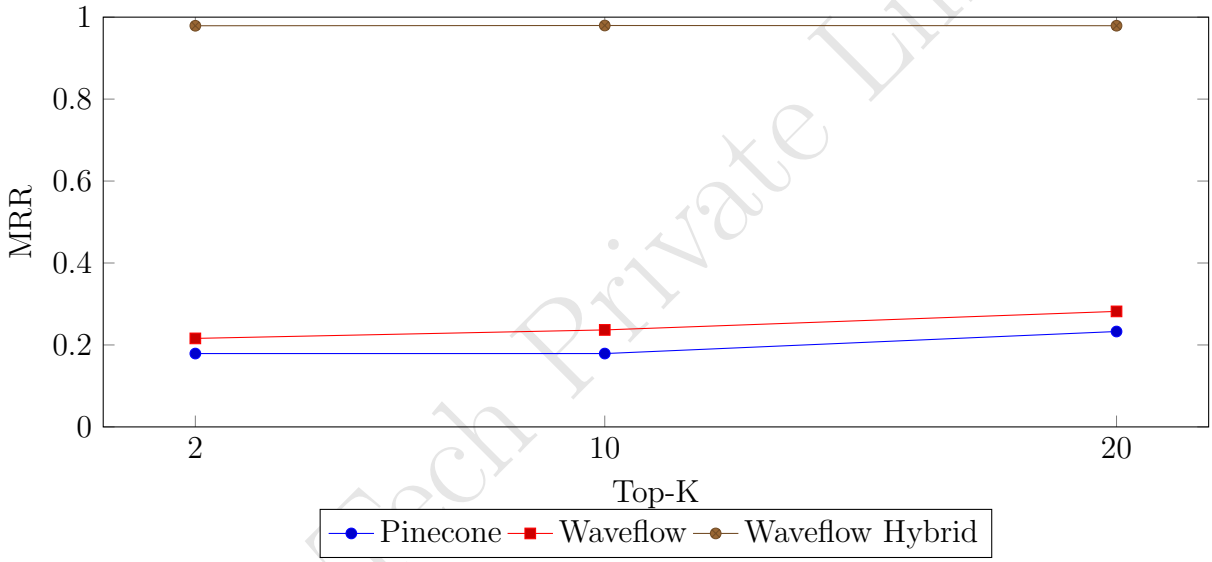

Figure 4: MRR trends across Top-K.

both precision and recall.

MRR and nDCG: Pinecone lags behind (< 0.25 MRR, < 0.16 nDCG). Waveflow dense-only improves moderately (MRR ~ 0.28, nDCG ~ 0.20). Hybrid consistently achieves near-perfect MRR (~ 0.98) and nDCG (~ 0.98), indicating strong ranking quality.

Efficiency: Pinecone benefits from faster queries at Top-K=2, but incurs embedding overhead. Waveflow eliminates embedding time, but has slightly higher query latency. Hybrid Waveflow adds negligible cost while delivering superior retrieval quality.

Overall: These results extend *Beyond Embedding Collapse* [1] by showing that hybrid retrieval strategies are not only theoretically sound but also practically efficient. System design choices directly impact collapse mitigation, ranking quality, and latency trade-offs.

4 Conclusion

This study demonstrates that embedding collapse is not just a theoretical concern but observable in production-scale benchmarks. Waveflow's hybrid architecture provides a viable remedy, achieving

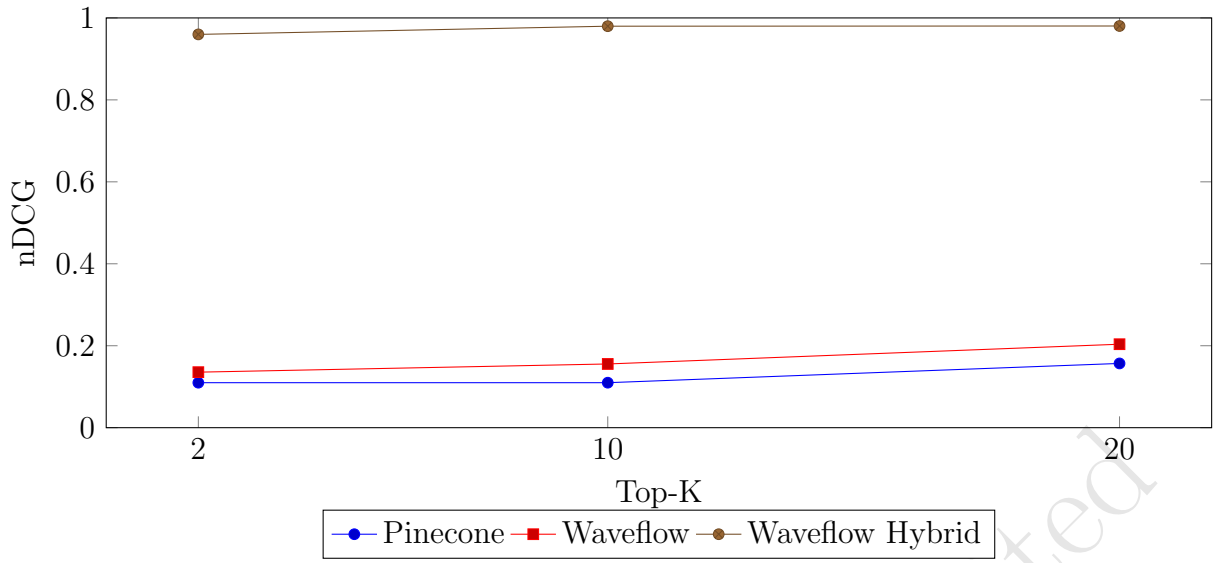

Figure 5: nDCG trends across Top-K.

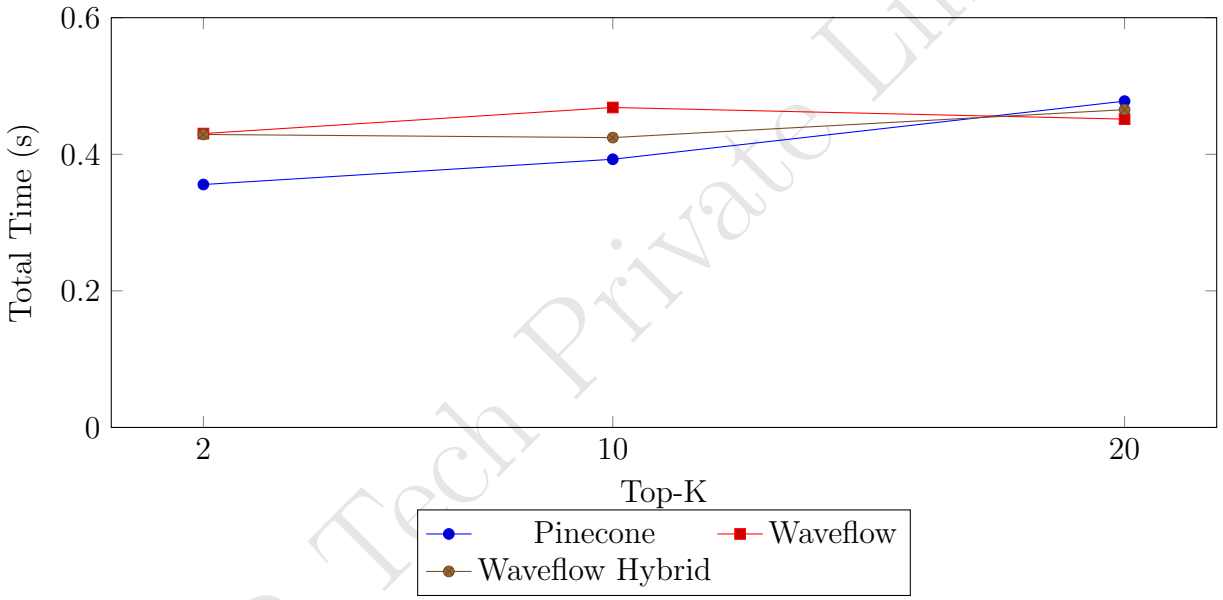

Figure 6: Latency comparison across Top-K.

high recall and ranking quality. For practitioners, this suggests that careful system design and hybridization strategies can mitigate the inherent limitations of dense-only embeddings.

References

[1] Beyond Embedding Collapse. *AlphaXiv Preprint*, 2025. Available at: `https://www.alphaxiv.org/pdf/2508.21038`