A Comparative Study

Alex Chen Priya Nair Marcus Webb
doany.ai Research Lab, San Francisco, CA alex.chen@doany.ai

Abstract

We evaluate retrieval latency across four strategies in a production RAG pipeline serving **12M monthly queries**. Our hybrid approach combining ColBERTv2 dense retrieval with BM25 sparse pre-filtering achieves **p95 latency of 47 ms** while maintaining **recall@10 of 0.94**—a **3.2× latency reduction** over naïve dense retrieval without sacrificing relevance.

Key Result

	Dense-only	Hybrid-seq
p95 Latency	152 ms	**47 ms**
Recall@10	0.91	**0.94**
Throughput	420 QPS	**1,100 QPS**

3.2× faster | **+3% recall** | **2.6× throughput**

Methods

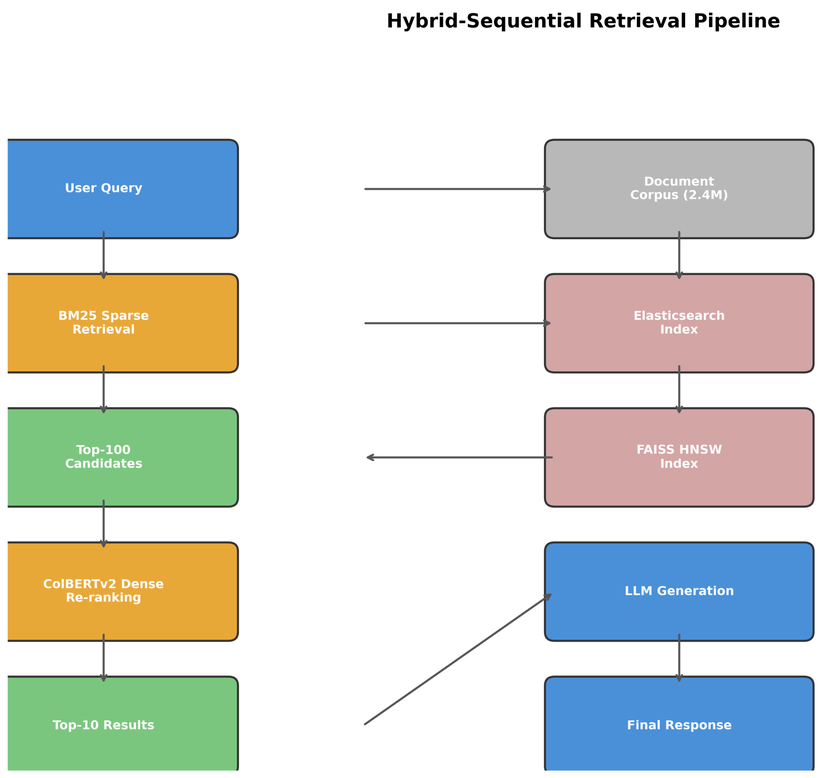
Retrieval Strategies Evaluated

1. **Dense-only**: ColBERTv2 over FAISS HNSW (768-dim)
2. **Sparse-only**: BM25 via Elasticsearch 8.x
3. **Hybrid-sequential**: BM25 top-100 → ColBERTv2 re-rank
4. **Hybrid-parallel**: Dense + Sparse with reciprocal rank fusion

Evaluation Setup

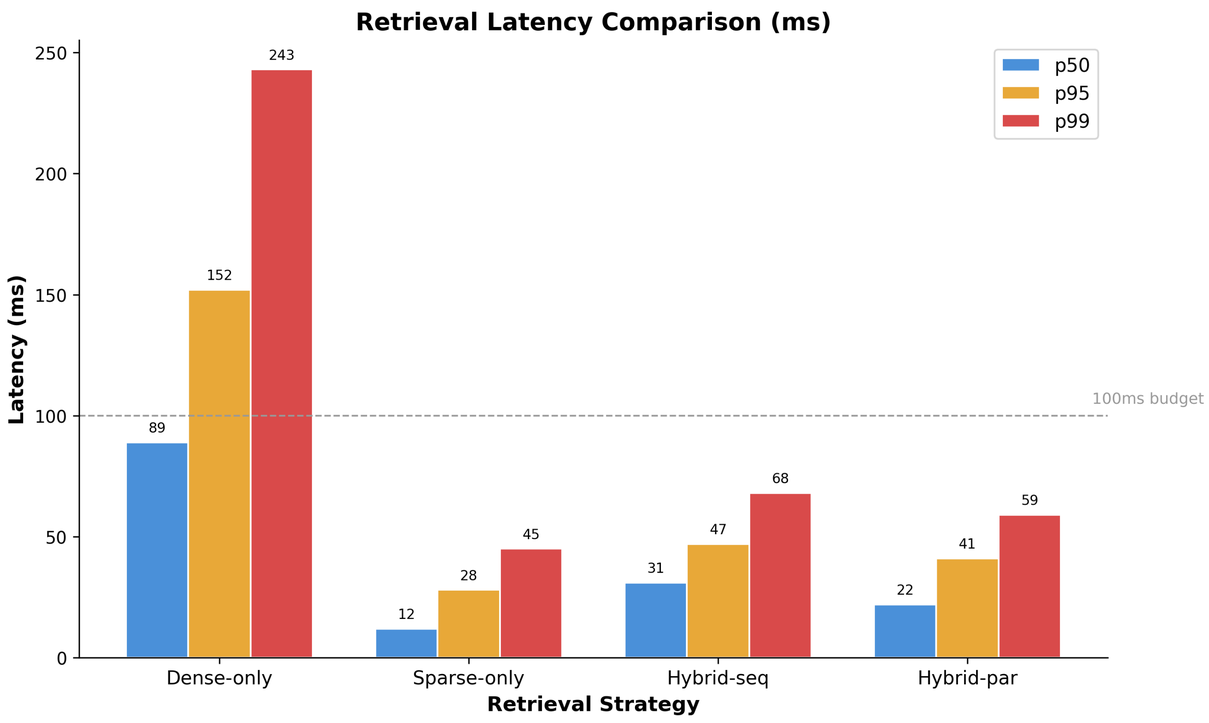
- Corpus: 2.4M skill documents (avg 1,200 tokens)
- 15,000 production queries (14 days)
- 4× NVIDIA A100 (80 GB) + 3-node ES cluster
- 3,200 human-annotated relevance judgments (Fleiss' $\kappa = 0.78$)

Pipeline Architecture



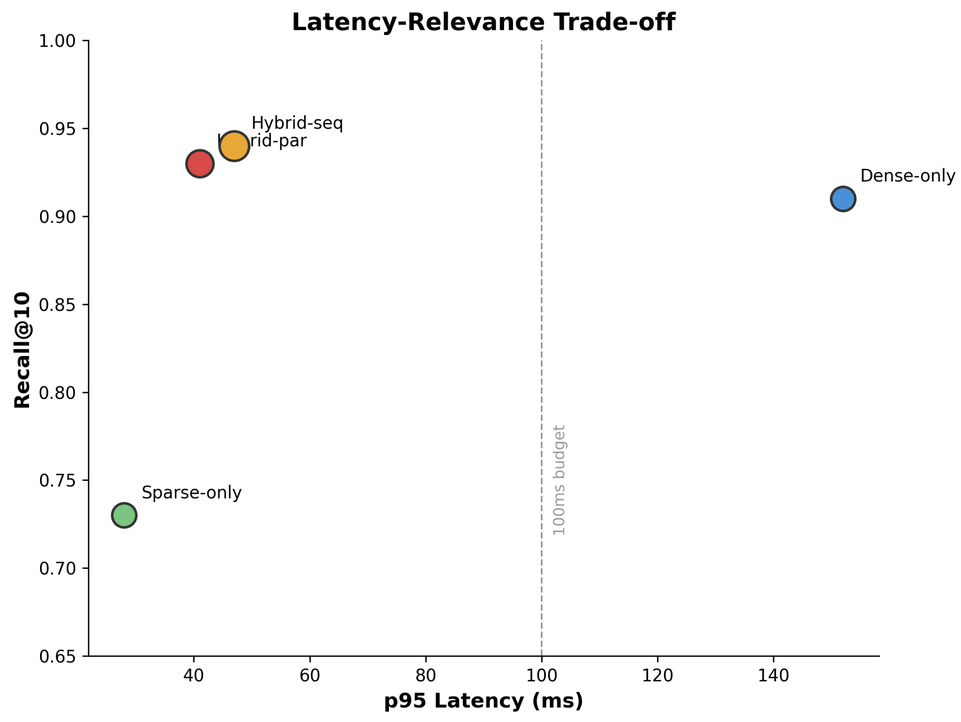
BM25 narrows 2.4M docs to top-100 in ∼12 ms; ColBERTv2 re-ranks in ∼18 ms.

Results: Latency Comparison



Strategy	p50	p95	p99	QPS
Dense-only	89	152	243	420
Sparse-only	12	28	45	2,800
Hybrid-sequential	31	**47**	68	1,100
Hybrid-parallel	22	41	59	1,450

Results: Latency–Relevance Trade-off



Strategy	Recall@10	nDCG@10	MRR
Dense-only	0.91	0.82	0.76
Sparse-only	0.73	0.65	0.58
Hybrid-seq	**0.94**	**0.87**	**0.83**
Hybrid-par	0.93	0.86	0.81

Conclusions

1. **Hybrid-sequential is optimal**: 3.2× faster than dense-only at p95 (47 ms vs 152 ms) with best recall@10 (0.94).
2. **BM25 pre-filtering is the key insight**: sparse retrieval as a first-pass filter reduces the dense search space

References

1. Khattab & Zaharia. *ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction.* SIGIR 2020.
2. Robertson & Zaragoza. *The Probabilistic Relevance Framework: BM25 and Beyond.* FnTIR 2009