Domain-Tuned Retrieval for RAG

Quantitative Improvements in Faithfulness and Answer Relevancy Under Shared Budgets

Alexander Talesnik
Knowledge2.ai
alex.talesnik@knowledge2.ai

Lev Muchnik
Knowledge2.ai
lev.muchnik@knowledge2.ai

Dave Lindon
Knowledge2.ai
dave.lindon@knowledge2.ai

September 19, 2025

Abstract

Modern AI systems, from retrieval-augmented generation (RAG) workflows to task-oriented agents, rely on high-quality retrieval over proprietary corpora. Recent theory ties embedding dimensionality to what a single-vector retriever can represent, motivating specialization over universal scaling [12]. Large, general-purpose embedding models provide a strong baseline but are often not aligned to domain-specific phrasing, abbreviations, and fine-grained distinctions. We study a targeted, implementation-agnostic approach: adapting compact retrievers to a given knowledge base using queries that reflect real user intent and a curriculum of hard negatives.

Under a **Shared Evaluation Budget**—Finance (SEC) top-K=5 with a cross-encoder reranker returning the top 3 passages, Clinical top-K=5 with no reranker, UniProtKB top-K=2 with no reranker, a shared GPT-4o-mini generator with max prompt tokens $\leq$ 3500, and identical AWS t3.large instances (2 vCPUs, 8 GB RAM)—we observe quantitative, absolute improvements over strong provider baselines. In the clinical domain ($N = 97$), our domain-tuned retriever improves Faithfulness by +0.0575 [95% BCa CI: 0.0203, 0.1120] vs OpenAI and +0.0454 [0.0135, 0.0957] vs Cohere, though the latter does not reach Holm-adjusted significance (p=0.0539); Answer Relevancy improves by +0.0696 [0.0418, 0.1241] vs OpenAI and +0.0771 [0.0464, 0.1366] vs Cohere. In Finance ($N = 97$), Faithfulness vs OpenAI shows a mean increase of +0.0641 that is not statistically significant (Holm-adjusted $p = 0.1652$), while answer relevance improves significantly by +0.0575 [0.0264, 0.1145] vs OpenAI and +0.0726 [0.0327, 0.1378] vs Cohere. In UniProtKB ($N = 97$), improvements are +0.1176 (Faithfulness) and +0.1358 (Answer Relevancy) vs OpenAI; +0.0774 (Faithfulness) and +0.0706 (Answer Relevancy) vs Cohere. Across domains, the largest gains are in **Answer Relevancy** alongside sharply reduced catastrophic error/abstention rates (Clinical: 5/97 vs 1/97; UniProtKB: 15/97 vs 1/97).

Finally, recent theory formalizes the limits of single-vector retrievers: for a fixed embedding dimension d, there exist top-k relevance patterns that cannot be represented, regardless of training data. Our results show that a *specialize-not-just-scale* strategy—compact, domain-tuned retrievers paired with a cross-encoder reranker—yields higher Faithfulness and Answer Relevancy on real corpora under shared budgets [12, 13], while reducing downstream hallucinations through more precise retrieval.

1 Introduction

Takeaway *Retrieval is the controllable lever in modern AI systems.* Whether the downstream component is a RAG pipeline or an agent acting in the world, answer quality is bounded by what is retrieved. Unlike prompt length—which increases Time-to-First-Token and is linked to higher hallucination and factual discrepancy rates—retrieval quality can be systematically engineered via indexing choices, query modeling, and contrastive training with a curriculum of hard negatives. Our focus is therefore on *making retrieval precise under fixed budgets*, not on ever-longer prompts or ever-larger general-purpose embedders.

Thesis *For domain-specific RAG, compact retrievers adapted to the target corpus through a curriculum of hard negatives can improve Faithfulness and Answer Relevancy relative to strong general-purpose baselines, under fixed computational budgets.*

This paper demonstrates that modest, targeted fine-tuning of compact models yields measurable retrieval improvements that translate directly to better downstream answer quality. We present a repeatable pipeline for domain adaptation and provide empirical evidence from three distinct domains: clinical trials, financial filings, and biological knowledge bases.

1.1 The Problem: Limits of General-Purpose Retrievers

Large models trained on general web data may underperform in specialized domains for several reasons:

1. **Distribution Shift.** Technical, financial, and scientific texts contain jargon, formal syntax, and abbreviations that diverge from the distribution of general web text.
2. **Query–Document Mismatch.** Users in specialized fields ask questions using shorthand and local vernacular, while source documents use formal language. Bridging this semantic gap requires domain-specific training.
3. **Negative Sampling Matters.** While modern baselines use in-batch and mined hard negatives [1, 6], domain-specific "near-miss" negatives—passages that are lexically or semantically similar but incorrect—are essential for learning fine-grained distinctions.
4. **Operational Inefficiency.** A common fallback for weak retrievers is to concatenate many marginally relevant passages. This is computationally wasteful, increases Time-to-First-Token (TTFT), and can heighten hallucinations by distracting the generator [4].

A note on theoretical limits. Recent theory shows a lower bound linking the embedding dimension d and the number of top-k document combinations that can be represented by a single-vector retriever: for any fixed d, there exist relevance patterns that *cannot* be realized, regardless of training data or optimization. An empirical dataset (LIMIT) constructed from this theory demonstrates that even SoTA embedders underperform on simple instantiations of these patterns. *Implication:* scaling dimensions or training data alone will not guarantee universal retrieval. Practical systems should instead restrict the task distribution and adopt architectures that reintroduce higher-resolution interactions (e.g., late interaction or rerankers) alongside compact, domain-tuned dual encoders.[1]

2 Methodology: A Domain-Tuning Pipeline

Pipeline overview. We implement a repeatable pipeline that ingests raw documents, normalizes them, models queries, and iteratively retrains compact retrievers. The indexing branch handles dense+sparse construction, clustering, hard-negative mining, batching, evaluation, and serving. Dense and sparse indices are refreshed after each adaptation epoch to align with the updated embeddings.

2.1 Data Readiness and Query Modeling

We ingest documents in common formats and normalize them into clean, provenance-preserving chunks. We then model queries to reflect realistic user intent, augmenting observed questions with synthetic paraphrases to cover terminology variants.

2.2 Contrastive Training with a Hard-Negative Curriculum

The core of our method is constructing **(query, positive, hard_negative)** triplets for contrastive training. Hard negatives are passages that are semantically or lexically similar to the correct answer but are definitively wrong. A curriculum that progresses from easy to near-miss hard negatives forces the model to learn the nuanced boundaries of the domain. Each query is paired with multiple hard negatives per batch to increase discrimination pressure and stabilize learning.

[1] See DeepMind's LIMIT [12] for the bound and dataset; see also late-interaction evidence from ColBERT [13].

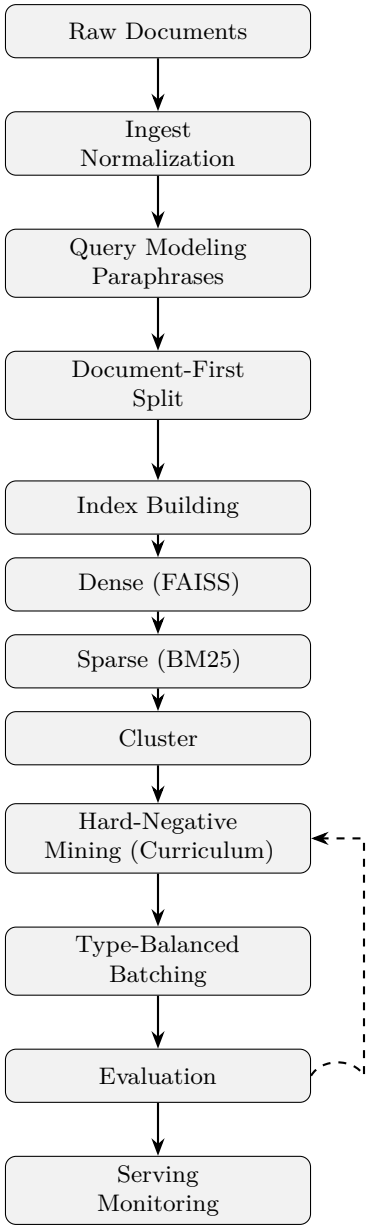

Figure 1: Domain-tuning pipeline. Evaluation feeds harder negatives into subsequent adaptation rounds while refreshed indices keep pace with the updated dual encoder.

Why this matters in practice. Dense retrievers are highly sensitive to negative sampling and topic overlap; near-miss negatives are essential to learn fine-grained decision boundaries. We use a staged curriculum (BM25/static → in-batch → mined hard) and refresh indices after each epoch to reduce false negatives and stabilize learning. This setup reflects best practice in the literature and addresses known brittleness of general-purpose dense models under distribution shift.[2]

Accessibility. Unlike prompt tuning for LLMs, retrieval adaptation requires dataset construction at scale (positives, mined near-misses, paraphrases), type-balanced batching, and continuous refresh. Without this, quality gains are often elusive [23, 24].

2.3 Two-Stage Retrieval for Serving

We adapt a compact dual-encoder retriever for fast "first-stage" retrieval. For enhanced precision, we optionally apply an off-the-shelf cross-encoder reranker for "second-stage" scoring on the top candidates.

[2]See Zhan et al. [23] and ANCE [6] on hard negatives; see also findings on update timing in evolving corpora [24].

This two-stage design balances latency and answer quality and mirrors evidence that late interaction or cross-encoder rerankers compensate for single-vector limits while keeping latency acceptable [13]. Throughout this paper, **ModernBERT ($\mathbf{K^2}$-tuned)** refers to the domain-specific retrievers produced by this pipeline using a ModernBERT-based architecture.

3 Evaluation Protocol

Style note. We capitalize metric names (**Faithfulness**, **Answer Relevancy**) consistently across text, tables, and figures. To isolate the retriever's contribution, we designed a rigorous, preregistered evaluation protocol.

Shared Evaluation Budget. All systems are compared under identical constraints using the RAGAS evaluation harness:

- Finance (SEC): retrieval top-K=5; cross-encoder reranker returns the top 3 passages to the generator.
- Clinical: retrieval top-K=5; single-stage retrieval (no reranker).
- UniProtKB: retrieval top-K=2; single-stage retrieval (no reranker).
- Generator: GPT-4o-mini (OpenAI) with temperature 0.0, nucleus sampling disabled, and maximum prompt length $\leq$ 3500 tokens.
- Hardware: AWS t3.large instances (2 vCPUs, 8 GB RAM) with CPU-only inference; batch size 1 for latency traces.
- Baselines: OpenAI `text-embedding-3-large`; Cohere `embed-v4.0`; configurations match provider defaults captured in the PAP.

Benchmarks vs. Business Value. Public leaderboards (e.g., MTEB, BEIR) provide breadth but also encourage general-purpose scaling. MTEB itself reports that no single embedding method dominates across tasks, while BEIR highlights out-of-distribution fragility and the enduring strength of BM25. Our evaluation therefore emphasizes *customer- and corpus-specific* end-user metrics under a *Shared Evaluation Budget*, rather than relying solely on leaderboard-style IR scores [19, 20].

Longer prompts increase TTFT without guaranteeing grounding; retrieval improves quality more efficiently [11].

3.1 Metrics and Terminology

We report the following end-user metrics, which are the primary focus of this study:

- **Faithfulness** (*document-grounding*): An answer is faithful if all factual claims are supported by the retrieved passages. Unsupported or contradictory claims reduce the score.
- **Answer Relevancy**: The degree to which the answer directly addresses the question, judged on a calibrated ordinal rubric and mapped to a [0,1] scale.

Faithfulness and Answer Relevancy are computed with RAGAS v1.2 using our domain-specific rubric prompts. These are model-based judgments (not human annotations). Uncertainty is estimated via a paired bootstrap over queries and we report 95% BCa confidence intervals (e.g., 10,000 reps; fixed seed).

3.2 Statistical Methods

We assess uncertainty with paired bootstrapping over queries: for each system comparison we draw 10,000 paired replicates (fixed random seed) and report 95% BCa confidence intervals on the mean difference. Significance is evaluated with a paired Wilcoxon signed-rank test (two-sided, $\alpha = 0.05$). Holm adjustment is applied per dataset $\times$ metric across the two baseline comparisons ($m = 2$), and adjusted p-values are reported alongside raw p-values in Appendix A.

3.3 Experimental Controls and Reproducibility

- **Pre-Analysis Plan (PAP):** We pre-specified our primary endpoints, hypotheses, statistical methods, and system configurations in a PAP to prevent p-hacking and ensure methodological rigor. The blinded PAP identifier is available under NDA.

- **Leakage Controls:** We apply document-first splitting and MinHash-based de-duplication to prevent train/test leakage.
- **Rater Protocol:** Faithfulness and Answer Relevancy scores are produced with RAGAS (v1.2) using adjudicated templates derived from our rubric.
- **Versioning:** All seeds, data splits, and run metadata are versioned for full reproducibility.

4 Empirical Results

We validated our approach on held-out test sets across three specialized domains: clinical trial documentation, financial filings, and the UniProtKB biological knowledge base. Each evaluation set contains $N = 97$ unique queries. Our adapted retrievers—ModernBERT models fine-tuned separately on each corpus by K^2 (*Knowledge Squared*)—were compared against strong baselines from OpenAI `text-embedding-3-large` and Cohere `embed-v4.0`.

Table 1 presents the primary results for Faithfulness and Answer Relevancy.

Table 1: Mean scores for Faithfulness and Answer Relevancy under the Shared Evaluation Budget. The domain-tuned retriever leads on average; the Finance Faithfulness lift vs OpenAI is not statistically significant (95% CI includes zero).

Dataset	System	Pipeline	Faithfulness	Answer Relevancy	N
Clinical	OpenAI	Dense-only	0.8505	0.8628	97
Clinical	Cohere	Dense-only	0.8625	0.8553	97
Clinical	**ModernBERT (K^2)**	**Dense-only**	**0.9079**	**0.9324**	97
Finance	OpenAI	Dense+Cross-Encoder Reranker	0.7189	0.8986	97
Finance	Cohere	Dense+Cross-Encoder Reranker	0.7449	0.8835	97
Finance	**ModernBERT (K^2)**	**Dense+Cross-Encoder Reranker**	**0.7830**	**0.9561**	97
UniProtKB	OpenAI	Dense-only	0.6037	0.8010	97
UniProtKB	Cohere	Dense-only	0.6439	0.8662	97
UniProtKB	**ModernBERT (K^2)**	**Dense-only**	**0.7213**	**0.9368**	97

Table 2 and Figure 2 quantify these improvements, showing the mean differences and 95% confidence intervals. Answer Relevancy gains are statistically significant across domains. Faithfulness gains are significant for Clinical vs OpenAI and for UniProtKB; the Clinical vs Cohere lift is treated as descriptive (Holm-adjusted p=0.0539), and Finance vs OpenAI shows a positive interval that does not meet the Holm-adjusted threshold (p=0.1652).

The domain-tuned model also reduces catastrophic errors and abstentions, a critical operational benefit. On the clinical set, ModernBERT (K^2-tuned) produces 1/97 catastrophic responses versus 5/97 for each baseline (Table 3); in UniProtKB, OpenAI abstains on 15/97 queries, a rate the adapted retriever cuts to 1/97 (Section 4.2).

4.1 Sensitivity to Catastrophic Errors

In many domains, failing to produce a supported answer is a critical failure. We define a "catastrophic" error as an unsupported answer or an abstention (e.g., "I don't know"). Table 3 shows that our adapted model produces far fewer catastrophic errors on the clinical dataset.

Extending the same diagnostic across corpora shows that the reduction in catastrophic outcomes persists (Table 4). Finance evaluations record a single catastrophic response for the adapted retriever versus 5/97 for OpenAI and 4/97 for Cohere; UniProtKB sees a 15/97 abstention rate for OpenAI that drops to 1/97 with the K^2 retriever.

Table 2: Per-domain improvements for ModernBERT (K^2-tuned): mean differences with 95% BCa CIs vs. baselines (N=97 per domain).

Dataset	Metric	vs. OpenAI	vs. Cohere
Clinical	Faithfulness	0.0575 [0.0203, 0.1120]	0.0454 [0.0135, 0.0957]
Clinical	Answer Relevancy	0.0696 [0.0418, 0.1241]	0.0771 [0.0464, 0.1366]
Finance	Faithfulness	0.0641 [0.0120, 0.1257]	0.0381 [−0.0201, 0.0996]
Finance	Answer Relevancy	0.0575 [0.0264, 0.1145]	0.0726 [0.0327, 0.1378]
UniProtKB	Faithfulness	0.1176 [0.0698, 0.1783]	0.0774 [0.0386, 0.1223]
UniProtKB	Answer Relevancy	0.1358 [0.0758, 0.2165]	0.0706 [0.0432, 0.1256]

Note: CIs are BCa on mean paired differences (10,000 bootstrap replicates; fixed seed). Significance uses the two-sided paired Wilcoxon signed-rank test with Holm correction per dataset×metric (m=2). Because the CI targets the mean difference while the test is rank-based, a CI that excludes 0 may still be marked "ns".

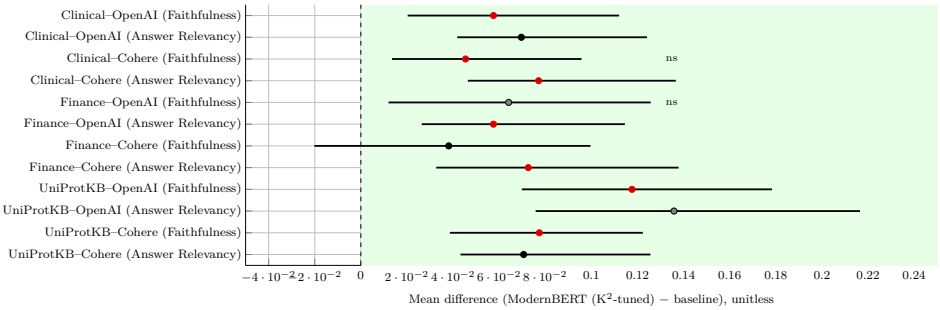

Figure 2: Forest plot of mean differences (ModernBERT − baseline). Markers denote means; whiskers denote 95% BCa CIs. Values to the right of the zero line (shaded region) indicate improvement. "ns" = not significant under two-sided paired Wilcoxon signed-rank with Holm correction ($\alpha = 0.05$); adjusted p-values in Appendix A.

Table 3: Sensitivity analysis (Clinical, $N = 97$). ModernBERT (K^2-tuned) leads under all treatments and exhibits a substantially lower catastrophic error rate.

	Faith. (mean)	Faith. (5% trim)	Faith. (non-zero)	Ans. Rel. (mean)	Catastrophic (#/%)
ModernBERT (K^2)	**0.9079**	**0.9296**	**0.9174**	**0.9324**	1 / 1.0%
Cohere	0.8625	0.8951	0.9094	0.8553	5 / 5.2%
OpenAI	0.8505	0.8820	0.8967	0.8628	5 / 5.2%

Table 4: Catastrophic error counts (Faithfulness=0) across datasets. Percentages are relative to 97 queries per set.

Dataset	System	Catastrophic (#/%)
Clinical	ModernBERT (K^2-tuned)	1 / 1.0%
Clinical	OpenAI	5 / 5.2%
Clinical	Cohere	5 / 5.2%
Finance	ModernBERT (K^2-tuned)	1 / 1.0%
Finance	OpenAI	5 / 5.2%
Finance	Cohere	4 / 4.1%
UniProtKB	ModernBERT (K^2-tuned)	1 / 1.0%
UniProtKB	OpenAI	15 / 15.5%
UniProtKB	Cohere	5 / 5.2%

4.2 Qualitative Analysis: Row-Level Debugging

Aggregate metrics are complemented by examining individual failures. In the UniProtKB domain, OpenAI abstained on 15/97 queries, whereas the adapted retriever reduces abstentions to 1/97. This pattern was consistent across domains and highlights the practical benefit of domain tuning. For example:

- **Clinical Query (Idx 22):** Both baselines abstained. Our model retrieved the correct combined-therapy context, enabling the generator to produce the supported description.
- **UniProtKB Query (Idx 19):** OpenAI abstained. Our model retrieved the passage detailing the protein's domain interaction, which the generator then summarized.

5 Discussion

5.1 A Practical Response to Theoretical Limits

LIMIT shows that single-vector models face dimension-dependent ceilings for representing arbitrary top-k relevance combinations. Our results indicate that, in applied settings, a combination of (i) compact, domain-tuned dual encoders and (ii) a second-stage cross-encoder reranker delivers higher Answer Relevancy and Faithfulness under shared budgets. This *specialize-not-just-scale* recipe narrows the query space to the domain's semantics and reintroduces higher-resolution comparisons where needed, mitigating the single-vector bottleneck without incurring large inference costs.[3]

5.2 Design Rationale: Why Domain Tuning Works

We attribute the observed gains to several factors:

1. **Targeted Supervision:** A moderate number of high-quality, domain-specific examples effectively shapes the embedding space toward relevant distinctions.
2. **Domain Alignment:** Fine-tuning reshapes embeddings to match the semantics of the target domain, teaching the model the specific meanings of jargon, acronyms, and symbols.
3. **Learning to Discriminate:** A staged hard-negative curriculum teaches fine-grained relevance boundaries and mitigates false-negative drag.
4. **RAG Efficiency:** Better retrieval leads to more concise, relevant prompts, which can reduce Time-to-First-Token (TTFT) and overall computational load.

5.3 Scale vs. Specialization

While provider embedding models are proprietary, it is reasonable to infer they are orders of magnitude larger than the compact ModernBERT family used here. Our results demonstrate that under a shared budget, specialization can be more effective than scale alone.

Table 5: Comparison of Specialized vs. General-Purpose Encoders

Dimension	ModernBERT (K^2-tuned)	General-purpose encoders
Accuracy (this study)	Higher within-domain scores on Faithfulness and Answer Relevancy under shared budgets.	Strong general-purpose performance; may lack precision on fine-grained domain tasks.
Latency & Cost	Compact encoder yields lower retrieval latency and cost, especially in on-prem/VPC deployments.	Larger encoders; latency and cost vary by SKU and deployment setup.
Deployment	Supports on-prem/VPC training and serving, offering data locality and privacy.	Often API-based; private options vary.
Adaptability	Versioned runs enable frequent, reproducible retraining as the corpus evolves.	Update cadence is opaque; domain adaptation options may be limited.

5.4 Operational Considerations

Beyond accuracy, our pipeline is designed for production environments, incorporating:

[3] As suggested by LIMIT, multi-vector or cross-encoder reranker stages can address cases that a single vector cannot represent; we adopt this in our rerank-K stage [12, 13].

- **Monitoring and Alerting:** Instrumentation for latency, error rates, and data drift.
- **Auditability and Governance:** Immutable run records, artifact-level provenance, and built-in support for PII/PHI redaction and secrets management.
- **Flexibility:** Compatible with common vector stores (FAISS, Elasticsearch, pgvector/PostgreSQL) and deployable in any cloud or on-prem environment.

6 Conclusion

For specialized RAG applications, domain-adapted compact retrievers can significantly improve Faithfulness and Answer Relevancy compared to strong, much larger, general-purpose alternatives under fixed budgets. The gains are most pronounced in domains with unique terminology and fine-grained semantic distinctions. Our end-to-end pipeline provides a reproducible, efficient, and operationally robust method for achieving these improvements.

Contribution Summary. We contribute a *retrieval adaptation framework* that: (1) ingests arbitrary corpora; (2) builds dense+sparse indices; (3) trains compact dual-encoders via a hard-negative curriculum with index refresh; (4) optionally adds a cross-encoder reranker; (5) evaluates under a shared budget with preregistered endpoints; and (6) *closes the loop* by capturing real interaction signals (queries and abstention flags) to mine harder negatives and auto-refresh models as the corpus and usage evolve. This yields accuracy and latency/cost competitive with much larger general models on in-domain tasks, while remaining deployable in VPC/on-prem environments [24].

Implications for Practitioners

- **For CTOs/VPs of Engineering:** Adapted compact retrievers can improve answer quality while simultaneously reducing latency and cost, especially in private cloud or on-prem deployments.
- **For ML Engineers:** The hard-negative curriculum is a powerful technique for reducing near-miss retrieval errors and improving grounding, which in turn enables more efficient use of the generator's context window.

Limitations & Scope. These results are demonstrated on three specific domains under the reported Shared Evaluation Budget. Performance may vary with different budgets, domains, or provider models. Potential threats to validity, such as query selection bias, are mitigated through our preregistered evaluation design.

Acknowledgments

We thank our colleagues at Knowledge2.ai for their feedback. All authors are affiliated with Knowledge2.ai.

Scope Note

This public whitepaper omits certain implementation specifics (e.g., prompts, hyperparameters). A more detailed technical brief and the preregistered PAP ID are available under NDA by contacting `research@knowledge2.ai`.

References

[1] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. *Proceedings of EMNLP*, 2020.

[2] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[3] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. Text and code embeddings by contrastive pre-training. *arXiv:2201.10005*, 2022.

[4] Xiaoqiang Lin, Aritra Ghosh, Bryan Kian Hsiang Low, Anshumali Shrivastava, and Vijai Mohan. REFRAG: Rethinking RAG-based Decoding. *arXiv:2509.01092*, 2025.

[5] Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrievers. *Proceedings of NAACL-HLT*, 2022.

[6] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *Proceedings of ICLR*, 2021.

[7] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Jingyuan Zhang, Jian-Yun Nie, and Ji-Rong Wen. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. *Proceedings of NAACL-HLT*, 2021.

[8] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *arXiv:2307.03172*, 2023. (TACL version 2024).

[9] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient Transformers: A Survey. *arXiv:2009.06732*, 2020.

[10] NVIDIA. LLM Inference Benchmarking: Fundamental Concepts (TTFT and throughput). Blog, 2025.

[11] Databricks. LLM Inference Performance Engineering: Best Practices (TTFT scaling with prompt length). Blog, 2023.

[12] Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. On the Theoretical Limitations of Embedding-Based Retrieval. *arXiv:2508.21038*, 2025. Data/code: https://github.com/google-deepmind/limit.

[13] Omar Khattab and Matei Zaharia. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. *Proceedings of SIGIR*, 2020.

[14] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised Dense Information Retrieval with Contrastive Learning. *Transactions on Machine Learning Research (TMLR)*, 2022.

[15] Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. Large Dual Encoders Are Generalizable Retrievers. *Proceedings of EMNLP*, 2022.

[16] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text Embeddings by Weakly-Supervised Contrastive Pre-training. *arXiv:2212.03533*, 2022.

[17] Rodrigo Nogueira and Kyunghyun Cho. Passage Re-ranking with BERT. *arXiv:1901.04085*, 2019.

[18] Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. Document Ranking with a Pretrained Sequence-to-Sequence Model (MonoT5). *Findings of EMNLP*, 2020.

[19] Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive Text Embedding Benchmark. *Proceedings of EACL*, 2023.

[20] Nandan Thakur, Nils Reimers, Andreas Ruckle, Abhishek Srivastava, and Iryna Gurevych. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. *NeurIPS Datasets and Benchmarks*, 2021.

[21] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *arXiv:1702.08734*, 2017.

[22] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval* 3(4), 2009.

[23] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. Optimizing Dense Retrieval Model Training with Hard Negatives. *Proceedings of SIGIR*, 2021.

[24] Dongyoung Ko, Jinhyuk Lee, and Minjoon Seo. When Should Dense Retrievers Be Updated in Evolving Corpora? *Findings of ACL*, 2025.

[25] Bradley Efron. Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association* 82(397), 1987.

[26] Frank Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* 1(6), 1945.

[27] Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6(2), 1979.

[28] Andrei Z. Broder. On the Resemblance and Containment of Documents. *Proceedings of Compression and Complexity of Sequences*, 1997.

[29] The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* 51(D1):D523–D531, 2023.

[30] U.S. National Library of Medicine. ClinicalTrials.gov. https://clinicaltrials.gov

[31] U.S. Securities and Exchange Commission. Accessing EDGAR Data. https://www.sec.gov/edgar/search/

A Significance Summary (Holm-Adjusted p-values)

Dataset	Baseline	Metric	p (Wilcoxon)	p (Holm)
Finance	OpenAI	Faithfulness	0.0826	0.1652
Finance	OpenAI	Answer Relevancy	9.846×10^{-6}	3.938×10^{-5}
Finance	Cohere	Faithfulness	0.2451	0.2451
Finance	Cohere	Answer Relevancy	0.001829	0.005487
Clinical	OpenAI	Faithfulness	0.0228	0.0456
Clinical	OpenAI	Answer Relevancy	$< 10^{-4}$	$< 10^{-4}$
Clinical	Cohere	Faithfulness	0.0539	0.0539
Clinical	Cohere	Answer Relevancy	$< 10^{-4}$	$< 10^{-4}$
UniProtKB	OpenAI	Faithfulness	1.637×10^{-4}	3.274×10^{-4}
UniProtKB	OpenAI	Answer Relevancy	2.804×10^{-8}	8.412×10^{-8}
UniProtKB	Cohere	Faithfulness	4.436×10^{-4}	4.436×10^{-4}
UniProtKB	Cohere	Answer Relevancy	4.441×10^{-14}	1.776×10^{-13}

B Dataset Cards (Summary)

- **Clinical.** Clinical trial documentation; sources and licenses recorded; redactions applied where required. Provenance retained per chunk. Detailed card available under NDA.
- **Finance.** SEC filings (e.g., 10-K); sources and licenses recorded; sensitive fields redacted. Provenance retained per chunk. Detailed card available under NDA.
- **UniProtKB.** Biological knowledge base; subset description, licenses, and redactions recorded. Provenance retained per chunk. Detailed card available under NDA.