A Representation Sharpening Framework for Zero Shot Dense Retrieval

Dhananjay Ashok[1]* **Suraj Nair**[2] **Mutasem Al-Darabsah**[2]
Choon Hui Teo[2] **Tarun Agarwal**[2] **Jonathan May**[2]
[1]Information Sciences Institute, University of Southern California [2]Amazon
ashokd@isi.edu {srjnair, mutasema, choonhui, tagar, jnatmay}@amazon.com

Abstract

Zero-shot dense retrieval is a challenging setting where a document corpus is provided without relevant queries, necessitating a reliance on pretrained dense retrievers (DRs). However, since these DRs are not trained on the target corpus, they struggle to represent semantic differences between similar documents. To address this failing, we introduce a training-free **representation sharpening** framework that augments a document's representation with information that helps differentiate it from similar documents in the corpus. On over twenty datasets spanning multiple languages, the representation sharpening framework proves consistently superior to traditional retrieval, setting a new state-of-the-art on the BRIGHT benchmark. We show that representation sharpening is compatible with prior approaches to zero-shot dense retrieval and consistently improves their performance. Finally, we address the performance-cost tradeoff presented by our framework and devise an indexing-time approximation that preserves the majority of our performance gains over traditional retrieval, yet suffers no additional inference-time cost.

1 Introduction

Dense retrieval systems represent queries and documents in a semantic space, using the similarity of their embeddings as an estimate of a document's relevance to a query (Yih et al., 2011). This approach has been shown to work well in data-rich domains like Question Answering (Karpukhin et al., 2020) and Web Search (Mao et al., 2022). However, in more realistic scenarios, we do not have access to relevant queries for the documents in our corpus (Dai et al., 2023) and must perform zero-shot retrieval (Gao et al., 2023). Approaches to this setting rely on generic, pretrained DRs, making alignment to the target corpus a major concern (Thakur et al., 2021; Wang et al., 2023). Such alignment often requires creating synthetic datasets for finetuning (Ma et al., 2021), incentivizing the DR to create document embeddings that are similar to those of relevant queries. While often successful (Izacard et al., 2022), retraining risks catastrophic forgetting (Goodfellow et al., 2014), can be computationally infeasible (Zhang et al., 2025), and, in the case of closed-source DRs (Neelakantan et al., 2022; Anthropic, 2025) is impossible. Training-free adaptation may avoid these limitations.

We introduce a framework of **representation sharpening** (Figure 1), that augments a document's embedding during inference to enhance a DR without retraining. To achieve this, we first generate synthetic queries either using standard query generation (Ma et al., 2021) or, with our novel **contrastive query generation** paradigm, which creates queries that contain information on what distinguishes a document from other, similar documents in the target corpus. During inference, we use the test query to assign weights to each generated query, and mix them into the document embeddings before computing the relevance score. This accentuates the unique aspects of each document, sharpening its representation with respect to the test query and enabling more precise retrieval.

On six datasets from the BEIR benchmark (Thakur et al., 2021), representation sharpening outperforms traditional inference and a document expansion baseline, improving NDCG@10 by 6.9% on average. We apply our framework to prior approaches to training-free, zero-shot dense retrieval, and show that it consistently improves all existing methods. Beyond standard benchmarks, we apply our framework to the complex reasoning-based BRIGHT (Su et al., 2025) benchmark, boosting pre-trained DRs and achieving a new state-of-the-art on eight of eleven measured subsets. We further show that our method is applicable to multiple languages, and improves the performance of

*Work done during an internship at Amazon.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3735–3751

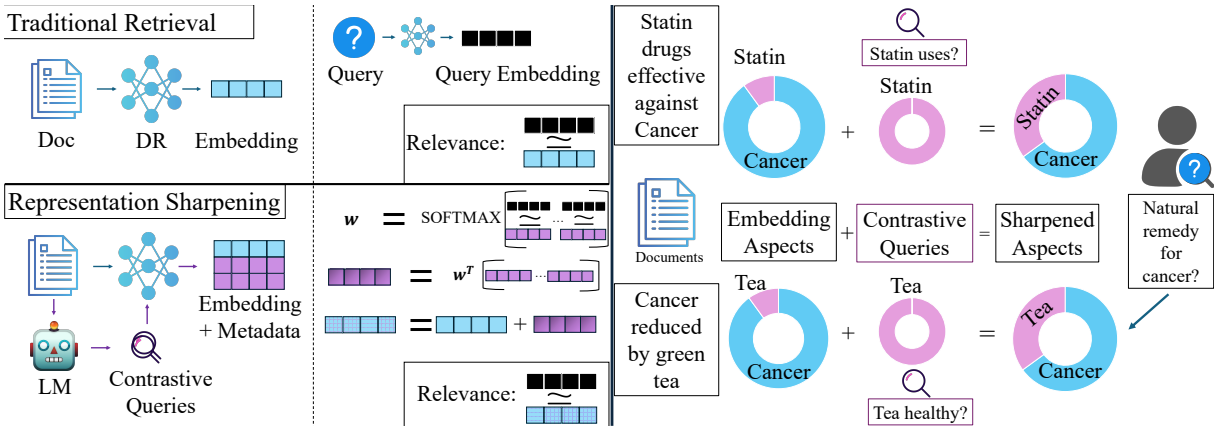

Figure 1: **Left**: Comparison of the traditional dense retrieval pipeline and our proposed framework for **representation sharpening**. During the indexing phase, we generate contrastive queries and store them along with a document's embedding as metadata. During inference, we leverage the user query to determine importance weights for each contrastive query. We then combine the contrastive query embeddings into the document embedding, sharpening its representation before computing retrieval relevance. **Right**: Example of how sharpening improves retrieval between two similar documents. Both documents discuss cancer, and that concept is more strongly encoded in their hidden representations (leftmost pie) than the unique aspects (such as 'statins'). We generate contrastive queries which focus on the unique aspects (representation in the middle pie). When we mix the query representations into the document embeddings, the result (rightmost pie) is a document representation that is no longer purely dominated by the 'cancer' aspect, making it easier for a test query to identify the correct document, as the document embeddings now more clearly encode the difference between the two.

DRs on the Hindi, Swahili, Korean and Thai splits of the MIRACL benchmark (Zhang et al., 2023).

Finally, we address implementation costs and present an alternative approach that, instead of adapting the inference procedure, performs a one-time alteration of the vector embeddings of the documents in the index. This indexing-time approach proves consistently superior to traditional retrieval and a document expansion baseline, showing that contrastive queries can boost retrieval while incurring **no additional inference-time cost**.

2 Background and Related Work

Statistical approaches to document indexing and retrieval have a long history (Bookstein and Swanson, 1974; Ponte, 1998), with most work assuming access to plentiful training data in the form of document-query relevance annotations (Bajaj et al., 2018; Kwiatkowski et al., 2019). However, while it has long been acknowledged that such a setting is limiting (Croft and Harper, 1979), performing retrieval zero-shot, i.e., without relevance labels was a historically challenging problem (Harman, 1992; Thakur et al., 2021). Modern methods tuned DRs on large training corpora and performed Zero-Shot transfer between domains (Yu et al., 2022; Wang et al., 2022; Lin et al., 2023). This approach relies on the target domain being well aligned with the training set (Chandradevan et al., 2024), which, as pointed out by Izacard et al. (2022) and Gao et al. (2023), can rarely be assumed, sparking an interest in alternate approaches (Dai et al., 2023).

The emergence of powerful LMs (Raffel et al., 2020) with impressive few-shot (Brown et al., 2020) and instruction-following (Ouyang et al., 2022) capabilities presented such an alternative. Significant gains have been made by leveraging these LMs to generate synthetic queries (Ma et al., 2021; Dai et al., 2023) and documents (Ma et al., 2023; Li et al., 2024) that can be used to train target-domain specific DRs. However, not only is adjusting the weights of DRs an often computationally restrictive process (Zhang et al., 2025), it is also prone to removing useful, pre-existing knowledge (Goodfellow et al., 2014). Additionally, for increasingly superior closed-source embedding systems (Neelakantan et al., 2022; Anthropic, 2025), weights are inaccessible, making fine-tuning impossible.

These concerns have prompted the development of training-free approaches (Gao et al., 2023) to enhancing pretrained DRs on unseen test domains. Recent methods (Wang et al., 2023; Shen et al., 2024) enhance the capabilities of DRs by expanding the inference queries to include more explicit information. However, this requires continuous

access to an LM during the inference phase, a prohibitively expensive design decision (Wu et al., 2022). In this work, we take an alternate approach and propose a solution that only requires LM access during the indexing phase of retrieval, increasing performance without relevance labels, retraining or significantly increasing deployment costs.

Query Generation: Synthetic query generation has proven a powerful approach to creating training sets for DRs (Ma et al., 2021; Sachan et al., 2022; Bonifacio et al., 2022). Subsequent work has focused on scaling the technique (Wang et al., 2024b), improving filtering (Almeida and Matos, 2024), incorporating domain knowledge (Xia et al., 2025) and capturing a range of possible user intents (Lee et al., 2024). Despite this progress, there has been little change to the fundamentally one-to-many paradigm of query generation, i.e., prior work generates queries that are relevant (or not relevant (Lv and Zhai, 2015)) to a single document in the corpus. In this work, we question this practice, proposing instead a **many-to-many** paradigm, that generates contrastive queries which are relevant to some documents, and simultaneously not relevant to other, similar documents in the corpus. This paradigm is orthogonal to specific improvements in query generation practices (e.g., incorporating additional information (Gupta et al., 2025)), as these methods can easily be applied within our paradigm as well. To the best of our knowledge, the only other works to explore such a paradigm do so for evaluation (Weller et al., 2024) and cross-language training dataset construction (Mayfield et al., 2023), while we are the first to show that such a framework for query generation can be used to actively improve performance without retraining.

We summarize our contributions as:

1. A zero-shot framework for **representation sharpening**, that boosts pretrained DRs on a wide range of datasets and languages.

2. **Contrastive query generation**, which creates queries that accentuate the aspects that most distinguish a document from other, similar documents within corpus.

3. An indexing-time method that directly alters the document embeddings in the index, and boosts retrieval performance while incurring **no additional inference-time cost**.

3 Representation Sharpening Framework

In this section, we first introduce the traditional paradigm of inference for dense retrieval and then provide a high-level overview of our alternative framework for **representation sharpening**. Finally, we explain how we operationalize this framework using LMs for contrastive query generation.

3.1 Traditional Retrieval

Consider a text corpus of documents $D = \{d_1, d_2 \ldots\}$ that must be ranked for relevance with respect to an inference query with text q. The traditional retrieval pipeline starts by precomputing the index of document representations $\mathbf{D} = \{\mathbf{d}_1, \mathbf{d}_2, \ldots\}$, where $\mathbf{d}_i \in \mathbb{R}^m$ is the semantic representation of d_i under some pretrained DR with embedding size m. During inference, we compute the representation $\mathbf{q} \in \mathbb{R}^m$ of the query and rank the documents based on $\mathrm{s} : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$, a measure of the similarity between the query and document representations. For instance, when using the standard measure of cosine similarity, the score of document d is computed as $\mathrm{s}(\mathbf{q}, \mathbf{d}) = \frac{\mathbf{q} \cdot \mathbf{d}}{\|\mathbf{q}\| \|\mathbf{d}\|}$

When two documents have similar representations $\mathbf{d}_1, \mathbf{d}_2$, it is usually the case that if query representation $\mathbf{q}$ is similar to $\mathbf{d}_1$, it will also be similar to $\mathbf{d}_2$. In such cases, if traditional retrieval inference is to determine which of the documents is **more** relevant, the representations must properly encode the subtle semantic differences between the two documents. However, this is often not the case for pretrained DRs (Ren et al., 2023).

3.2 Representation Sharpening

We propose that instead of exclusively relying on the DR to represent these differences, we explicitly provide this information in the form of **contrastive queries**. Given a document d, the set of contrastive queries Q_d are the queries for which d is relevant, and some other, similar documents in the corpus are **not** relevant. We compute their representations $\mathbf{Q}_d = \{\mathbf{q}_{d,1}, \mathbf{q}_{d,2} \ldots\}$ and store them as meta-data for document d in the index. During inference, we shift the document embedding in the direction of the contrastive query embeddings with:

$$\mathbf{d}^* = \mathbf{d} + \alpha \cdot g(\mathbf{q}, \mathbf{Q}_d) \quad (1)$$

where g is an inference-query aware aggregation function over the contrastive query embeddings. In our work, we consider g to be a convex combination of the contrastive query embeddings, where

DR	Variant	FiQA	NFCorpus	SciFact	T-COV	SciDocs	Arguana	Avg
Contriever	Trad	24.19	31.23	57.05	23.54	14.38	49.43	33.30
	Doc2Query	20.14	31.71	57.20	28.47	14.40	50.29	33.70
	SimSharp (Ours)	29.45	32.81	65.61	33.25	17.24	53.75	38.69
	ConSharp (Ours)	**30.21**	**34.89**	**68.51**	**35.36**	**17.94**	**57.06**	**40.67**
Qwen3	Trad	34.28	30.41	64.61	58.49	16.48	66.60	45.15
	Doc2Query	35.29	30.83	64.65	**68.52**	20.06	65.02	47.40
	SimSharp (Ours)	38.90	33.52	70.51	65.26	21.78	66.18	49.35
	ConSharp (Ours)	**40.62**	**34.30**	**71.88**	62.90	**22.55**	**69.97**	**50.37**
E5-Mistral	Trad	36.89	25.75	64.79	44.66	2.69	50.16	37.49
	Doc2Query	34.62	20.07	64.89	39.81	6.36	47.59	35.55
	SimSharp (Ours)	43.08	33.78	72.23	**50.73**	7.07	57.03	43.98
	ConSharp (Ours)	**45.15**	**35.51**	**74.33**	50.25	**8.57**	**59.42**	**45.54**

Table 1: NDCG@10 on BEIR benchmark when comparing traditional retrieval (Trad) to our proposed representation sharpening approach (SimSharp and ConSharp). Across datasets and underlying dense retrievers, ConSharp outperforms both traditional inference and a Doc2Query baseline with an average improvement of 6.9%.

the weights are determined by how similar they are to the inference query, as judged by a softmax:

$$g(\mathbf{q}, \mathbf{Q}_d) = \sum_{\mathbf{q}_{d,i} \in \mathbf{Q}_d} \frac{\exp(\mathrm{s}(\mathbf{q}, \mathbf{q}_{d,i}))}{\sum_{\mathbf{q}_{d,j} \in \mathbf{Q}_d} \exp(\mathrm{s}(\mathbf{q}, \mathbf{q}_{d,j}))} \mathbf{q}_{d,i} \tag{2}$$

This 'sharpens' (Berthelot et al., 2019; Huang et al., 2025) the document's representation, and reinforces the aspects that most saliently distinguish it from the other documents in the corpus in a way that is relevant to the inference query at hand. The relevance score of d is then $\mathrm{s}(\mathbf{q}, \mathbf{d}^*)$.

3.3 Selecting Contrastive References

To facilitate the framework described above, given a document d, we must generate contrastive queries that differentiate between d and other, similar documents in the corpus. To do so, we must select the similar documents that will serve as *contrastive references* for d. A natural choice is to leverage the DR and select the nearest neighbors of d in the document index $\mathbf{D}$. However, a document may simultaneously address a variety of topics, and selecting the nearest neighbors risks a failure to capture the full diversity of a document's topic range. Instead, we subsample a large, local neighborhood of d in $\mathbf{D}$ and identify groups using unsupervised clustering algorithms (Aggarwal et al., 1999), e.g., KMeans (Jin and Han, 2010). We then select one document d' from each cluster to form our set of contrastive references $D'_d = (d'_1, d'_2, \ldots)$.

3.4 Generating Contrastive Queries

For a given document d and set of contrastive references D'_d, we instruct an LM to create queries that are relevant to d, but irrelevant to some $(d'_i, d'_j, \ldots) \subseteq D'_d$. This is a departure from existing approaches to query generation (Ma et al., 2021; Sachan et al., 2022), that perform one-to-many generation, i.e., generating queries for one document at a time. Under the standard paradigm, it is unclear whether a relevant query generated for d is also relevant to d', and can produce generic queries that do not accentuate the unique contents of d. For instance, given either document in Figure 1 (right), standard query generation may produce 'How to prevent cancer?'. However, this is generic, as it is applicable to both documents. We instead propose a **many-to-many** paradigm for query generation that allows us to create more precise queries that highlight unique document aspects. As shown in Section 4, these precise queries prove more capable of separating a document from other, similar documents in the corpus.

4 Outperforming Traditional Inference

We use six datasets from the BEIR benchmark, spanning the domains of financial QA (FiQA), scientific understanding (SciDocs, SciFACT, NFCorpus), scientific news (Trec-COVID) and counter argument mining (Arguana). We select some of the most widely used pretrained DRs: Contriever (Izacard et al., 2022), Qwen3Embedding-0.6B (Zhang

Method	Inference	FiQA	NFCorpus	SciFact	Trec-COVID	SciDocs	Arguana	Avg
HyDE	Traditional	37.88	34.47	73.86	86.41	21.90	55.63	51.69
	ConSharp	**42.18**	**36.18**	**78.20**	**86.51**	**22.61**	**57.02**	**53.78**
Query2Doc	Traditional	41.04	34.97	74.70	86.62	22.37	65.70	54.23
	ConSharp	**44.06**	**36.94**	**78.09**	**87.51**	**23.42**	**68.38**	**56.40**
LameR	Traditional	39.29	38.55	73.85	83.14	21.96	50.63	51.24
	ConSharp	**42.85**	**39.44**	**76.94**	**84.82**	**22.82**	**54.29**	**53.53**
ReDE-RF	Traditional	36.47	36.44	68.24	**82.94**	20.96	58.36	50.56
	ConSharp	**38.94**	**37.93**	**72.22**	82.63	**22.95**	**60.93**	**52.60**

Table 2: NDCG@10 when prior methods are deployed under traditional inference v.s. the representation sharpening framework with contrastive queries (ConSharp). ConSharp outperforms on all datasets, improving average NDCG@10 by 2.2% and showing its consistent ability to boost the performance of prior methods.

et al., 2025) and E5-Mistral (Wang et al., 2024a) and repeat the following experiment:

Contrastive Reference Selection: Given a corpus D and pretrained DR, we compute the document index $\mathbf{D}$. For each document in the index, we subsample the top 100 nearest neighbors (by cosine similarity) and then cluster them using KMeans. Following standard practices (Shahapure and Nicholas, 2020), we select the number of clusters $k^* \in [3, 10]$ which achieves the highest silhouette score $\sum_{i=1}^{100} \frac{b_i - a_i}{\max(a_i, b_i)}$, where a_i is the average distance of neighbor i to other points in the same cluster and b_i is the minimum average distance of neighbor i to other clusters. This allows us to select a variable number of clusters for each document based on the range of topics it covers. We select the neighboring document that is closest to the centroid of each cluster, and compile them to form the set of k^* contrastive references: D'_d.

Contrastive Query Generation: For each contrastive reference document $d'_i \in D'_d$, we use Claude-3.5-Sonnet (Anthropic, 2024) to generate contrastive queries that are relevant to d and not relevant to d'_i. We combine the queries generated across all contrastive references to form Q_d. We use the same prompt template for all datasets and do not engage in any prompt engineering (for the template and more details see Appendix D).

Representation Sharpening: During inference, we use Eq 1 with a fixed $\alpha = 1$, giving equal weight to the document and query components.

Baselines: We compare our method against the traditional retrieval pipeline. Additionally, we follow standard practice in query generation (Ma et al., 2021) and generate queries under a one-to-many paradigm i.e., from each individual document in the corpus (hereby called *simple queries*). These queries form a stronger baseline: *Doc2Query*, where the simple queries are concatenated to the end of the text of the document before embedding (Nogueira et al., 2019). We benchmark these baselines against two variants of our method — *SimSharp*, where the queries Q_d used to shift the document in Eq 1 are the simple queries, and *ConSharp*, where Q_d is the set of contrastive queries. We use the standard BEIR metric of NDCG@10, with alternate metrics reported in Appendix A.

Results: Across all datasets, and retrievers (Table 1), representation sharpening consistently outperforms all baselines, with ConSharp achieving an average NDCG@10 improvement of 6.9% over traditional inference. While both SimSharp and ConSharp provide significant gains, results show that using contrastive queries consistently boosts performance, validating the merit of our many-to-many query generation approach. This consistency underscores the generality of the method, and its ability to boost performance on a wide range of retrieval tasks, regardless of the underlying DR.

4.1 Improving Previous Methods

Recent work in zero-shot dense retrieval follows the pseudo-relevance feedback (PRF) paradigm (Li et al., 2023) and acts on the inference query q. For example, HyDE (Gao et al., 2023) and Query2Doc (Wang et al., 2023) prompt an LM to directly answer q and use this answer to obtain a refined query embedding $\mathbf{q}_r$. Follow-up work like LameR (Shen et al., 2024) advances answer generation, while ReDE-RF (Jedidi et al., 2024) uses real documents instead of generated answers, but all PRF methods maintain the fundamental approach

System	Stack Exchange							Coding		Theorem-Based	
	Bio	Earth	Econ	Psy	Rob	Stack	Sus	Leet	Pony	AoPS	TheoT
Grit-LM	25.0	32.8	19.0	19.9	17.3	11.6	18.0	29.8	**22.0**	8.8	21.1
OpenAI	23.7	26.3	20.0	27.5	12.9	12.5	20.3	23.6	2.5	8.5	12.3
Voyage	23.6	25.1	19.8	24.8	11.2	15.0	15.6	30.6	1.5	7.4	11.1
Google	23.0	**34.4**	19.5	27.9	16.0	17.9	17.3	29.6	3.6	9.3	14.3
Reason-IR-8B	26.2	31.4	23.3	**30.0**	18.0	23.9	20.5	35.0	10.5	14.7	27.2
+ ConSharp	**27.5**	31.5	**23.5**	29.9	**18.5**	**25.7**	**21.7**	**36.1**	10.8	**14.9**	**28.3**

Table 3: NDCG@10 on BRIGHT benchmark when representation sharpening is used on the ReasonIR-8B DR (+ ConSharp, other method results from Shao et al. (2025)). ConSharp consistently boosts performance, leading to the ReasonIR-8B + ConSharp system establishing a new state-of-the-art on eight of eleven possible splits.

DR	Variant	Ko	Th	Hi	Sw
MCont	Trad	29.31	40.73	14.58	32.25
	Doc2Q	21.07	18.18	8.54	33.59
	S.Sharp	39.89	**52.91**	23.75	**47.04**
	C.Sharp	**40.16**	51.89	**24.41**	46.62
E5-M	Trad	47.05	61.82	30.71	61.34
	Doc2Q	44.87	56.27	30.65	59.72
	S.Sharp	**54.82**	**67.83**	40.61	65.64
	C.Sharp	52.29	65.72	**41.35**	**66.76**

Table 4: NDCG@10 on MIRACL datasets. Representation sharpening consistently improves performance, showing that the method can boost dense embedders on tasks that span across languages.

of obtaining a richer $\mathbf{q}_r$ and computing document relevance with $\mathrm{s}(\mathbf{q}_r, \mathbf{d})$ instead. While these solutions operate on the query at inference time, our approach augments the document's representation. This makes our approach compatible with PRF, as changing $g(\mathbf{q}, \mathbf{Q}_d)$ in Eq 1 to $g(\mathbf{q}_r, \mathbf{Q}_d)$ combines representation sharpening with these methods. We use Qwen3-Embedding-0.6B as a dense retriever, and measure the performance of these methods when using traditional inference vs. ConSharp.

Results: When used in conjunction with previous methods, representation sharpening consistently outperforms traditional inference (Table 2), raising average NDCG@10 by 2.2%. This shows that the framework synergizes well with prior work and can be seamlessly included to boost performance.

4.2 State-of-the-art on retrieval for reasoning

BEIR consists of information-seeking queries where DRs perform well. However, complex queries may require reasoning that goes beyond surface form matching (Su et al., 2025). The BRIGHT benchmark consists of such queries, with domains such as arithmetic and coding. We use ReasonIR-8B (Shao et al., 2025), a DR which achieves state-of-the-art performance on BRIGHT, and measure the performance increase when using representation sharpening by contrastive queries (ConSharp). We set $\alpha = 0.2$, based on a hyperparameter search in the range $\alpha \in [0.05, 1.5]$ on one of BRIGHT's subsets (TheoremQA Questions, arbitrarily chosen) and omit this subset from the results.

Results: Representation sharpening consistently boosts ReasonIR on BRIGHT (Table 3). When combined, the new system achieves state-of-the-art performance on eight of eleven measured subsets. This shows that the framework can be helpful even in complicated domains like reasoning or coding.

4.3 Boosting multilingual retrievers

Representation sharpening leads to improvements on a range of English-language datasets. However, given that LMs are known to perform better on English than on other languages (Asai et al., 2024), it is unclear whether the method can perform well in the multilingual case. To test this, we select four diverse low-resource languages from the MIRACL benchmark (Korean, Hindi, Swahili, Thai) and use the multilingual DRs of MContriever (Izacard et al., 2022) and E5-Multilingual-Large-Instruct (Wang et al., 2024c). Since Claude-3.5-Sonnet has shown impressive multilingual capabilities (Enis and Hopkins, 2024), we do not change the generating model. We use the same English prompt template from previous experiments, as the LM proves capable of generating contrastive queries in the appropriate language for each document.

Results: Regardless of the language or DR, representation sharpening offers consistent and considerable performance boosts (Table 4), with ConSharp

improving average NDCG@10 by 8.9%. Surprisingly, while sharpening with simple queries (SimpSharp) delivers fewer gains on the BEIR datasets (Table 1), it performs competitively in the multilingual setting, suggesting that LMs like Claude-3.5-Sonnet lack the power to create nuanced contrastive queries in non-English languages.

5 Index-Time Sharpening

The representation sharpening framework provides consistent performance gains on a wide variety of domains and languages, however it comes at a cost. Each document is paired with a set of contrastive queries Q_d, the embeddings of which are stored in the index as meta-data. This leads to the index growing by a factor of $\frac{1}{|D|}\sum_{d\in D}|Q_d|$. Seeking to eliminate any inference time cost to the algorithm, we identify strategies by which we may use contrastive queries to augment a document's representation during the indexing phase alone.

The first strategy is document expansion (DocExp) (Nogueira et al., 2019); for every document d, we generate contrastive queries $Q_d = \{q_{d,1}, q_{d,2}, \ldots\}$ and expand the document by concatenating the text of the queries: $d_{\text{expand}} = d \circ q_{d,1} \circ q_{d,2} \ldots$. The index then stores the embedding of this expanded document $\mathbf{d}_{\text{expand}}$ and the relevance score is given by $\text{s}(\mathbf{q}, \mathbf{d}_{\text{expand}})$.

We further introduce a novel method of more directly editing the index, *IndexSharp*. We take inspiration from the parameter-free adapter (PEFA) (Chang et al., 2024), which selects training set queries and uses them to augment the document's representation in the index, and extend the method to the zero-shot setting. We set g in Eq 1 to $\frac{1}{|Q_d|}\sum_{\mathbf{q}_d\in\mathbf{Q}_d}\mathbf{q}_d$, removing dependence on inference queries and isolating cost to indexing. We implement both of these alternatives, using Qwen3-Embedding-0.6B as the underlying DR.

Results: All approaches consistently outperform the traditional retrieval pipeline (Table 5), with IndexSharp improving average NDCG@10 by 4.7%. While ConSharp method (inference-time) is typically superior, it is encouraging to see that the performance gap is not wide, suggesting that the bulk of our performance improvements can be preserved with no inference time cost.

6 Ablations and Analysis

We conduct ablations and analysis with the goal of better understanding which components and design decisions most contribute to the method's success.

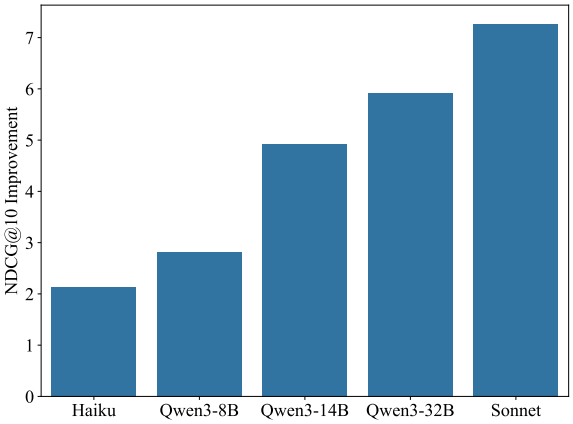

Figure 2: Performance improvement on SciFact when varying the underlying LM. Performance scales with model size, however Qwen3-8B still delivers a noteably performance boost, showing that even open-weight models on the 8B parameter scale can be used to generate effective contrastive queries.

Importance of Document Quality: For each test query, we retrieve the top 100 documents and compute each document's effective *sharpening boost*: $\text{s}(\mathbf{q}, \mathbf{d}^*) - \text{s}(\mathbf{q}, \mathbf{d})$. A high sharpening boost implies that using representation sharpening leads to the relevance score increasing by a significant amount. We then follow prior work (Dathathri et al., 2020) and compute the perplexity as a measure of document fluency, and observe the Pearson correlation (Pearson, 1895) between these two values. Across all datasets, we see a negative correlation (Table 6), suggesting that the more fluent a document is, the higher its sharpening boost. The result stresses the importance of crafting documents that are of high quality, as this may effect the boost they receive.

Language Model: With the SciFact dataset and the Qwen3-Embedding-0.6B DR, we vary the LM that generates the contrastive queries, using Claude-3-Haiku and open-weight models from the Qwen3 (Zhang et al., 2025) family. The results (Figure 2) show that performance scales with the size of the model. However, even Qwen3-8B gives a significant performance boost, showing that representation sharpening is not reliant on the largest of models, and can be useful even when powered by smaller, more accessible open weight models. Encouragingly, the performance gap between the Qwen3-32B model and Claude-3-5-Sonnet is only 1.35%, suggesting that a sufficiently powerful open-source model can enable similar performance gains as the most capable frontier models.

Variant	FiQA	NFCorpus	SciFact	Trec-COVID	SciDocs	Arguana	Avg
Traditional	34.28	30.41	64.61	58.49	16.48	66.60	45.15
DocExp	34.81	30.85	64.65	**71.32**	20.14	64.95	47.78
IndexSharp	**40.74**	<u>33.80</u>	<u>69.73</u>	62.73	<u>22.45</u>	<u>69.60</u>	<u>49.84</u>
ConSharp	<u>40.62</u>	**34.30**	**71.88**	<u>62.90</u>	**22.55**	**69.97**	**50.37**

Table 5: NDCG@10 of ConSharp when compared (**best**, <u>second-best</u>) to approaches that offload all cost to the indexing operation. Indexing time sharpening (IndexSharp) consistently outperforms other index-time baselines, and captures the majority of ConSharp's performance increases while incurring no additional inference-time costs.

DR	FiQA	NFCorpus	SciFact	Trec-COVID	SciDocs	Arguana	Avg
Contriever	-2.36	-5.81	-13.71	-6.77	-2.19	-6.19	-6.18
Qwen3	-1.88	-13.85	-8.45	-3.81	-4.94	-11.75	-7.45
E5-Mistral	-4.29	-3.52	-8.72	-9.35	-2.86	-9.36	-6.35

Table 6: Pearson correlation between the perplexity and sharpening boost received by documents. More fluent documents receive greater boosts, highlighting the important of creating high quality and fluent documents.

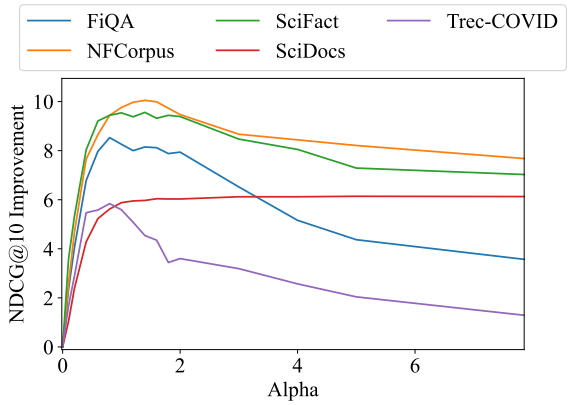

Figure 3: Performance when varying α on E5-Mistral. Increasing α from 0 to 1 always leads to improvements and performance declines after peaking in the $\alpha \in [1, 1.5]$ range. The best value of α is never the value used in our experiments ($\alpha = 1$), suggesting that hyperparameter tuning can provide further gains.

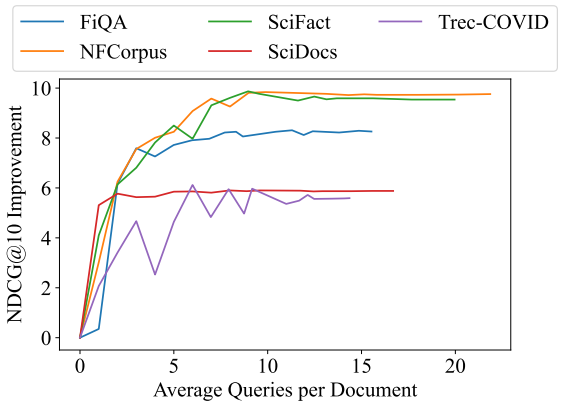

Figure 4: Performance when varying the average number of queries used per document on E5-Mistral. Using more queries increases performance, with the majority of the improvement occurs with the first 10 queries.

Hyperparameters: Our framework consists of hyperparameter α, that controls the weight between the document and the query component of the sharpened representation, and an implicit n, i.e., the number of contrastive queries used (per document) during inference. We fix the DR to the E5-Mistral model and vary these two. We see (Figure 3) that increasing α from 0 to 1.5 leads to significant improvements in performance, after which performance declines. The peak performance for each dataset is never achieved by $\alpha = 1$, which is the value we use in our experiments. This shows that, should a validation set be available, representation sharpening could be further improved by tuning the value of α. Increasing the number of queries used (Figure 4) is always beneficial. With $n > 10$, performance begins to plateau, suggesting that 10 queries are enough for the system to reach high levels of performance.

7 Conclusion

We tackle the challenging setting of zero-shot dense retrieval, and propose a framework for **representation sharpening** that boosts the performance of a DR without retraining. To operationalize our framework, we introduce a **many-to-many** paradigm for query generation, deploying it to create **contrastive queries** that accentuate the aspects that distinguish

a document from other, similar documents in the corpus. Through extensive experimentation, we show that representation sharpening outperforms traditional retrieval inference on a variety of tasks and languages, boosts prior methods and sets a new state-of-the-art on the BRIGHT benchmark. Finally, we devise an indexing time algorithm that achieves considerable performance gains, showing that representation sharpening can boost performance with **no additional inference-time cost**.

8 Limitations

Much like prior work in zero-shot dense retrieval (Gao et al., 2023; Wang et al., 2023), the representation sharpening framework relies on Language Models, using them for query generation. This imposes a natural limitation on the range of domains where the method can be expected to deliver gains, based on the capability of the LM used. In domains where the LM suffers from a lack of comprehension (an unknown language or niche and technical context), we would not expect our method to perform well. Additionally, as is common with methods in training-free adaptation (Gao et al., 2023; Chang et al., 2024), while we are able to improve performance, the extent of the improvement does not match that of training a superior DR. This can be seen in Table 1, where, though ConSharp outperforms the traditional baseline on the Contriever model, it is generally not sufficient to make the Contriever model better than the Qwen3-Embedding-0.6 model. This suggests that while the method can offer significant performance gains, it cannot replace a superior underlying retrieval system. Finally, while we limit our scope to training-free approaches, contrastive query generation naturally produces contrastive triplets of the form (Q_d, d, d') and future work may explore the merits of using such queries for contrastive learning (Izacard et al., 2022).

9 Ethical Considerations

This work leverages LMs for synthetic data generation, raising several important ethical considerations. Recent studies have found LMs to contain a variety of implicit biases on axes spanning gender, race, political affiliation, etc. (Resnik, 2025). Practitioners must consider their specific choice of generating model, and investigate how potential biases could affect the queries generated. Additionally, the most powerful LMs are often closed-source, API-access models (Anthropic, 2024; Brown et al., 2020), as opposed to open-weight models that can be run on internal hardware. To use the representation sharpening framework with these models, one must send data through API calls, a procedure which could expose confidential data to external actors. To ensure an equitable use of the method, practitioners must consider the privacy protocols of the inference API service under consideration, and ask whether the corpus used for query generation contains sensitive information.

References

Charu C. Aggarwal, Stephen C. Gates, and Philip S. Yu. 1999. On the merits of building categorization systems by supervised clustering. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, page 352–356, New York, NY, USA. Association for Computing Machinery.

Tiago Almeida and Sérgio Matos. 2024. Exploring efficient zero-shot synthetic dataset generation for information retrieval. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1214–1231, St. Julian's, Malta. Association for Computational Linguistics.

Anthropic. 2024. Claude Haiku 3.5 — anthropic.com. https://www.anthropic.com/claude/haiku. [Accessed 22-07-2025].

Anthropic. 2025. Embeddings - Anthropic — docs.anthropic.com. https://docs.anthropic.com/en/docs/build-with-claude/embeddings. [Accessed 30-07-2025].

Akari Asai, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2024. BUFFET: Benchmarking large language models for few-shot cross-lingual transfer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1771–1800, Mexico City, Mexico. Association for Computational Linguistics.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A human generated MAchine Reading COmprehension dataset. *Preprint*, arXiv:1611.09268.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. 2019. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. InPars: Unsupervised dataset generation for information retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2387–2392, New York, NY, USA. Association for Computing Machinery.

Abraham Bookstein and Don R Swanson. 1974. Probabilistic models for automatic indexing. *Journal of the American Society for Information Science*, 25(5):312–316.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *European Conference on Information Retrieval*, volume 9626, pages 716–722. Springer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ramraj Chandradevan, Kaustubh Dhole, and Eugene Agichtein. 2024. DUQGen: Effective unsupervised domain adaptation of neural rankers by diversifying synthetic query generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7437–7451, Mexico City, Mexico. Association for Computational Linguistics.

Wei-Cheng Chang, Jyun-Yu Jiang, Jiong Zhang, Mutasem Al-Darabsah, Choon Hui Teo, Cho-Jui Hsieh, Hsiang-Fu Yu, and S. V. N. Vishwanathan. 2024. PEFA: Parameter-free adapters for large-scale embedding-based retrieval models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 77–86, New York, NY, USA. Association for Computing Machinery.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

W.B. Croft and D.J. Harper. 1979. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35(4):285–295.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith Hall, and Ming-Wei Chang. 2023. Promptagator: Few-shot dense retrieval from 8 examples. In *The Eleventh International Conference on Learning Representations*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

Maxim Enis and Mark Hopkins. 2024. From llm to nmt: Advancing low-resource machine translation with claude. *Preprint*, arXiv:2404.13813.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

Ian J. Goodfellow, Mehdi Mirza, Xia Da, Aaron C. Courville, and Yoshua Bengio. 2014. An empirical investigation of catastrophic forgeting in gradient-based neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Nitin Gupta, Manish Kesarwani, Sambit Ghosh, Sameep Mehta, Carlos Eberhardt, and Dan Debrunner. 2025. Schema and natural language aware in-context learning for improved GraphQL query generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 1009–1015, Albuquerque, New Mexico. Association for Computational Linguistics.

Donna K. Harman. 1992. Relevance feedback revisited. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Audrey Huang, Adam Block, Dylan J Foster, Dhruv Rohatgi, Cyril Zhang, Max Simchowitz, Jordan T. Ash, and Akshay Krishnamurthy. 2025. Self-improvement in language models: The sharpening mechanism. In *The Thirteenth International Conference on Learning Representations*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Nour Jedidi, Yung-Sung Chuang, Leslie Shing, and James Glass. 2024. Zero-shot dense retrieval with embeddings from relevance feedback. *Preprint*, arXiv:2410.21242.

Xin Jin and Jiawei Han. 2010. *K-Means Clustering*, pages 563–564. Springer US, Boston, MA.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and

Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, et al. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Yoonsang Lee, Minsoo Kim, and Seung-won Hwang. 2024. Disentangling questions from query generation for task-adaptive retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4775–4785, Miami, Florida, USA. Association for Computational Linguistics.

Hang Li, Ahmed Mourad, Shengyao Zhuang, Bevan Koopman, and Guido Zuccon. 2023. Pseudo relevance feedback with deep language models and dense retrievers: Successes and pitfalls. *ACM Transactions on Information Systems*, 41(3):1–40.

Xiaopeng Li, Xiangyang Li, Hao Zhang, Zhaocheng Du, Pengyue Jia, Yichao Wang, Xiangyu Zhao, Huifeng Guo, and Ruiming Tang. 2024. Syneg: Llm-driven synthetic hard-negatives for dense retrieval. *Preprint*, arXiv:2412.17250.

Sheng-Chieh Lin, Amin Ahmad, and Jimmy Lin. 2023. mAggretriever: A simple yet effective approach to zero-shot multilingual dense retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11688–11696, Singapore. Association for Computational Linguistics.

Yuanhua Lv and ChengXiang Zhai. 2015. Negative query generation: bridging the gap between query likelihood retrieval models and relevance. *Inf. Retr.*, 18(4):359–378.

Guangyuan Ma, Xing Wu, Peng Wang, Zijia Lin, and Songlin Hu. 2023. Pre-training with large language model-based document expansion for dense passage retrieval. *Preprint*, arXiv:2308.08285.

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2021. Zero-shot neural passage retrieval via domain-targeted synthetic question generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1075–1088, Online. Association for Computational Linguistics.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836.

Kelong Mao, Zhicheng Dou, Hongjin Qian, Fengran Mo, Xiaohua Cheng, and Zhao Cao. 2022. ConvTrans: Transforming web search sessions for conversational dense retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2935–2946, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

James Mayfield, Eugene Yang, Dawn Lawrie, Samuel Barham, Orion Weller, Marc Mason, Suraj Nair, and Scott Miller. 2023. Synthetic cross-language information retrieval training data. *Preprint*, arXiv:2305.00331.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, et al. 2022. Text and code embeddings by contrastive pre-training. *Preprint*, arXiv:2201.10005.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *Preprint*, arXiv:1904.08375.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Karl Pearson. 1895. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242.

Jay Michael Ponte. 1998. *A language modeling approach to information retrieval*. Ph.D. thesis, University of Massachusetts Amherst.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Ruiyang Ren, Yingqi Qu, Jing Liu, Xin Zhao, Qifei Wu, Yuchen Ding, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2023. A thorough examination on zero-shot dense retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15783–15796, Singapore. Association for Computational Linguistics.

Philip Resnik. 2025. Large language models are biased because they are large language models. *Computational Linguistics*, 51(3):885–906.

Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R. Hersh. 2021. Searching for scientific evidence in a pandemic: An overview of trec-covid. *J. of Biomedical Informatics*, 121(C).

Devendra Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3781–3797, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ketan Rajshekhar Shahapure and Charles Nicholas. 2020. Cluster quality analysis using silhouette score. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 747–748.

Rulin Shao, Rui Qiao, Varsha Kishore, Niklas Muennighoff, Xi Victoria Lin, Daniela Rus, Bryan Kian Hsiang Low, Sewon Min, Wen-tau Yih, Pang Wei Koh, and Luke Zettlemoyer. 2025. Reasonir: Training retrievers for reasoning tasks. *arXiv preprint arXiv:2504.20595*.

Tao Shen, Guodong Long, Xiubo Geng, Chongyang Tao, Yibin Lei, Tianyi Zhou, Michael Blumenstein, and Daxin Jiang. 2024. Retrieval-augmented retrieval: Large language models are strong zero-shot retriever. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15933–15946, Bangkok, Thailand. Association for Computational Linguistics.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han yu Wang, Liu Haisu, Quan Shi, Zachary S Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O Arik, Danqi Chen, and Tao Yu. 2025. BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. In *The Thirteenth International Conference on Learning Representations*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia. Association for Computational Linguistics.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2345–2360, Seattle, United States. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024a. Text embeddings by weakly-supervised contrastive pre-training. *Preprint*, arXiv:2212.03533.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024c. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Orion Weller, Dawn Lawrie, and Benjamin Van Durme. 2024. NevIR: Negation in neural information retrieval. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2274–2287, St. Julian's, Malta. Association for Computational Linguistics.

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, et al. 2022. Sustainable ai: Environmental implications, challenges and opportunities. In *Proceedings of Machine Learning and Systems*, volume 4, pages 795–813.

Yu Xia, Junda Wu, Sungchul Kim, Tong Yu, Ryan A. Rossi, Haoliang Wang, and Julian McAuley. 2025. Knowledge-aware query expansion with large language models for textual and relational retrieval. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4275–4286, Albuquerque, New Mexico. Association for Computational Linguistics.

Wen-tau Yih, Kristina Toutanova, John C. Platt, and Christopher Meek. 2011. Learning discriminative

projections for text similarity measures. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 247–256, Portland, Oregon, USA. Association for Computational Linguistics.

Yue Yu, Chenyan Xiong, Si Sun, Chao Zhang, and Arnold Overwijk. 2022. COCO-DR: Combating the distribution shift in zero-shot dense retrieval with contrastive and distributionally robust learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

A Alternate Metrics

Aside from the traditional metric of NDCG@10, we also provide measurements of Recall@50 (Table 3) and MAP@50 (Table 8), to show that the consistent outperformance of representation sharpening is robust across metrics.

B Accounting of Cost Over Pipeline:

We provide a full account of the indexing time cost for the method below. We start with a corpus embedding index (organised in an HNSW index (Malkov and Yashunin, 2018)) and perform the following 4 steps.

1. Corpus Neighbour Computation: We use Approximate K-Nearest Neighbours to identify the 100 most similar neighbour documents for each document in the corpus. This is a $O(\log n)$ operation repeated n times, hence $O(n \log n)$. Note that exact retrieval is neither required nor preferred in this operation, as we merely seek to identify similar documents to choose as contrastive references.

2. Contrastive Reference Document Selection: Next, for each document D, we use KMeans on the 100 neighbours to identify representative contrastive references. This operation is $O(1)$ for each document (as the factors which determine the KMeans complexity i.e., number of datapoints to cluster, etc. are constant) and runs n times i.e: $O(n)$

3. Prompting: Next, for each document and contrastive reference pair, we compute contrastive queries with a single prompted LM forward pass. This can be made more efficient by combining all contrastive references into the same prompt, hence making us query the LM a total of n times.

4. Embedding Queries: Finally, we embed the contrastive queries generated (an $O(n)$ operation overall) and store them as metadata along with the document embeddings.

5. In total, our procedure is an $O(n \log n)$ operation, with n separate calls to a LM. This is of the same order as the construction of the HNSW index itself, which is an $O(n \log n)$ operation and is part of most efficient retrieval stacks already. Note that if we construct the HNSW index with this method in mind, we are able to save the document-to-document similarity scores and perform both operations at once.

We believe this procedure adds a manageable, log-linear cost to the indexing phase, while providing significant gains over the standard retrieval paradigm. However, for those who wish to reduce costs further, there are optimisations they could consider.

- Stage 1: Our method can operate on specific subsets of the document base, while leaving the remaining portions unaffected. Practitioners can leverage prior work on document selection to filter the documents whose representations they wish to sharpen. This would greatly reduce the operation cost from $O(n \log n)$ to $O(n_s \log n_s)$ where n_s is the size of the subset of documents being sharpened.

- Stage 2: The Contrastive Reference Document Selection can be avoided by selecting the top 3 neighbours directly. We observe that this reduces performance; however, the system still improves over standard retrieval.

- Stage 3: As shown in Fig 2, the usage of smaller, open-source models still enables superior performance to standard retrieval. For

DR	Variant	FiQA	NFCorpus	SciFact	Trec-COVID	SciDocs	Arguana
Contriever	Trad	46.40	23.34	86.70	1.59	29.15	94.24
	Doc2Query	40.95	23.52	86.70	2.31	29.36	95.31
	SimSharp	**51.09**	24.52	89.93	1.81	31.62	96.23
	ConSharp	**51.83**	**24.77**	**90.17**	**1.97**	**32.35**	**96.51**
Qwen3	Trad	60.67	24.48	88.89	6.04	36.07	98.51
	Doc2Query	59.64	24.34	88.56	6.18	38.30	98.22
	SimSharp	63.04	**25.68**	90.67	**6.68**	40.71	98.58
	ConSharp	**64.32**	**26.55**	**90.89**	6.62	**41.14**	**98.86**
E5-Mistral	Trad	59.58	23.49	89.00	3.36	6.50	94.24
	Doc2Query	58.86	20.18	88.67	**14.01**	3.39	93.74
	SimSharp	64.89	25.79	90.83	3.66	8.99	95.80
	ConSharp	**66.95**	**26.64**	**91.50**	3.75	**9.05**	**95.92**

Table 7: Recall@50 on BEIR benchmark. Across datasets and underlying dense retrievers, representation sharpening outperforms traditional inference.

DR	Variant	FiQA	NFCorpus	SciFact	Trec-COVID	SciDocs	Arguana
Contriever	Trad	19.92	14.18	53.29	0.93	9.71	42.10
	Doc2Query	16.19	14.46	53.49	**1.64**	9.73	42.55
	SimSharp	**23.98**	15.01	61.25	1.27	11.53	45.95
	ConSharp	23.91	**15.91**	**64.06**	1.35	**12.10**	**49.31**
Qwen3	Trad	28.53	13.45	60.48	6.65	11.37	58.98
	Doc2Query	29.53	13.69	60.35	7.30	14.01	57.45
	SimSharp	32.34	15.03	65.8	**7.30**	15.04	58.91
	ConSharp	**34.70**	**15.52**	**66.05**	7.13	**15.67**	**62.89**
E5-Mistral	Trad	31.11	10.95	60.58	2.64	1.64	43.08
	Doc2Query	29.32	8.51	60.71	**3.97**	2.58	40.61
	SimSharp	36.70	15.05	67.92	2.93	4.17	48.84
	ConSharp	**38.84**	**16.22**	**70.52**	3.02	**5.22**	**49.41**

Table 8: MAP@50 on BEIR benchmark. Across datasets and underlying dense retrievers, representation sharpening outperforms traditional inference.

those who have access to computational resources and wish to avoid API costs, they may leverage large-scale open-source models (30B parameters or more) to achieve substantial performance boosts while avoiding financial costs.

C Illustrative Example:

We walk through our pipeline and showcase the mechanism by which representation sharpening improves retrieval. We begin with the two documents in our index that have been deemed similar.

```
Document 1: Statin use after diagnosis
of breast cancer and survival: a
population-based cohort study...

Document 2: Dietary intakes of mushrooms
and green tea combine to reduce the risk
of breast cancer in Chinese women. ....
```

We then generate contrastive queries, which are meant to be relevant to document 1, but not document 2:

```
1. What biological mechanisms explain
the association between statin use
and improved survival in breast cancer
patients?
2. What are the clinical implications
```

```
of prescribing statins to breast
cancer patients after diagnosis?
```

Note the comparison of these queries against simple queries that are generated for document 1, without any contrastive reference:

```
1. Did the study control for confounding
factors such as other medications,
lifestyle, or socioeconomic status?
2. What implications does this study
have for breast cancer patients?
3. Does tea use influence breast
cancer-specific survival compared to
non-users?
```

The list of simple queries is suboptimal, as some of these queries (1 and 2) apply to Document 2 as well. In contrast, the contrastive queries are more refined, focusing on the aspects that only Document 1 answers, and providing more useful information for sharpening. In particular, in its reasoning, the generating model identifies:

```
Since Document 1 focuses on statin
use after breast cancer diagnosis
and survival outcomes,
while Document 2
is about dietary prevention
(mushrooms and green tea) before
diagnosis, the queries must be about
post-diagnosis treatment and survival
using statins, not prevention or diet.
```

This refined focus allows us to generate queries that most accentuate the unique features of one document over another.

For another example, see the document pair below:

```
Document 1: Just have the associate
sign the back and then deposit it.
It's called a third party cheque
and is perfectly legal. I wouldn't
be surprised if it has a longer
hold period and, as always, you
don't get the money if the cheque
doesn't clear...

Document 2: Lets say you owed me
\$123.00 an wanted to mail me a check.
I would then take the check from my
mailbox an either take it to my
bank, or scan it and ...
```

Both documents involve reference to checks, however while the first document discuss depositing a check, the second discusses mailing a check. To accentuate the differences between document 1 and document 2, we would want contrastive queries that discuss the most unique aspects of document 1. This is one of the contrastive queries our system generates for this pair of documents:

```
How do you deposit a third-party
check at a bank?
```

As we can see, this query discusses the aspect that makes the document most unique (depositing, as opposed to a more generic query regarding a check or banks etc), and hence allows us to sharpen its representation and make it easier to differentiate from document 2.

D Contrastive Query Generation Prompts

Given a document text d, we are provided with a contrastive reference document d' and tasked with generating contrastive queries that are relevant to d but not relevant to d'. We achieve this using the following prompt:

```
Here are some examples of queries to
understand the style:
[EXAMPLE 1]
[EXAMPLE 2]
[EXAMPLE 3]
[EXAMPLE 4]
[EXAMPLE 5]

Given the following two documents,
create a query that is weakly
related to both documents such that
document 1 is directly relevant to
the query, but document 2 is not.
Your plan should highlight the key
difference between the documents
that you will use.

Document 1: [DOCUMENT d]
Document 2: [Document d']
Return as many answers as you can,
but make sure that each answer is
unique and distinct. Do not repeat
yourself across answers, and focus on
quality over quantity.
Format your answer in the following
output structure:

<PLAN>Explanation on how you will design
the query and why the first document is
```

relevant to it but the second is not.
Also explain how you will ensure the
style is similar to the style of the
queries provided above
(name the language you will use)</PLAN>

<QUERY>text of the query in the same
language as the examples and
document above</QUERY>

We use the **same** 5 example queries for all (d, d') pairs of a given dataset, ensuring our method does not require access to more than 5 queries from the domain of interest.

For the Arguana dataset alone, we ask for a 'counter-argument passage' instead of a 'query', keeping all other parts of the prompt the same.

We then parse the output text using regular expressions to extract the generated queries.

D.1 Example Outputs:

FiQA Dataset

Document 1: Just have the associate
sign the back and then deposit it.
It's called a third party cheque and
is perfectly legal. I wouldn't be
surprised if it has a longer hold
period and, as always, you don't get
the money if the cheque doesn't clear.
Now, you may have problems if it's
a large amount or you're not very
well known at the bank. In that
case you can have the associate
go to the bank and endorse it in
front of the teller with some ID.
You don't even technically have to
be there. Anybody can deposit money
to your account if they have the
account number. He could also just
deposit it in his account and write
a cheque to the business.

Document 2: "Lets say you owed me
$123.00 an wanted to mail me a
check. I would then take the check
from my mailbox an either take
it to my bank, or scan it and
deposit it via their electronic
interface. Prior to you mailing
it you would have no idea which
bank I would use, or what my
account number is. In fact I
could have multiple bank accounts,
so I could decide which one to
deposit it into depending on what
I wanted to do with the money,
or which bank paid the most
interest, or by coin flip. Now
once the check is deposited my
bank would then ""stamp"" the
check with their name, their
routing number, the date, an my
account number. Eventually an
image of the canceled check would
then end up back at your bank.
Which they would either send to you,
or make available to you via
their banking website. You don't
mail it to my bank. You mail it
to my home, or my business, or
wherever I tell you to mail it.
Some business give you the address
of another location, where either
a 3rd party processes all their
checks, or a central location
where all the money for multiple
branches are processed. If you
do owe a company they will generally
ask that in the memo section in
the lower left corner that you
include your customer number.
This is to make sure that if they
have multiple Juans the money
is accounted correctly. In all
my dealings will paying bills
and mailing checks I have never
been asked to send a check
directly to the bank. If they
want you to do exactly as you
describe, they should provide
you with a form or other instructions."

Contrastive Queries:
1. How do you deposit a third-party
check at a bank?
2. Is endorsing a check in front of a
teller necessary for deposit?
3. Can you deposit money into someone
else's account with just their
account number?

NFCorpus Dataset

Document 1: Statin use after diagnosis
of breast cancer and survival: a

```
population-based cohort study...

Document 2: Dietary intakes of mushrooms
and green tea combine to reduce the risk
of breast cancer in Chinese women. ....

Contrastive Queries:
1. statins breast cancer
2. simvastatin
3. Cohort studies breast cancer
```

E Datasets

We use the following datasets in our experiments. All datasets / splits used are available under permissive licenses that are in line with our use case.

BEIR Benchmark: From the BEIR benchmark (Thakur et al., 2021), we select all datasets that have under 200K documents in the corpus:

- FiQA (Maia et al., 2018): A financial question answering dataset with 57K documents and 648 test queries
- NFCorpus (Boteva et al., 2016): A dataset for biomedical IR with 3.6K documents and 323 test queries
- SciFACT (Wadden et al., 2022): A dataset for scientific fact verification with 5K documents and 300 test queries
- Trec-COVID (Roberts et al., 2021): A dataset for scientific retrieval of COVID related documents with 171K documents and 50 test queries
- SciDocs (Cohan et al., 2020): A dataset of with 25K documents and 1000 test queries
- Arguana (Wachsmuth et al., 2018): A dataset for counterargument mining with 8.67K documents and 1406 test queries.

MIRACL Benchmark:

We select four of the lowest resource languages from the MIRACL (Zhang et al., 2023) benchmark, ensuring to maintain diversity across regions and language families:

- Hindi: 148,107 documents 350 test queries
- Korean: 437,373 documents 213 test queries
- Swahili: 47,793 documents 482 test queries
- Thai: 128,179 documents 733 test queries

BRIGHT Benchmark:

We use all splits of the BRIGHT (Su et al., 2025) benchmark.

F Hardware

Experiments were run on two systems:

1. System 1: Most experiments. 32 CPUs, 244 GiB memory, Processor: Intel Xeon E5-2686 v4 (Broadwell), Clock Speed: 2.3 GHz and 4 NVIDIA-TeslaV100 GPUs each with 16GB GPU memory.
2. System 2: For only the most compute intensive jobs (Qwen3 LM ablation). 24 CPUs, 1152 GiB memory, Processor: Intel Xeon Platinum 8275L, Clock Speed: 3 GHz and 8 NVIDIA-A100 GPUs, each with 40GB GPU memory.