Negative Sampling Techniques in Information Retrieval: A Survey

Laurin Wischounig[*], **Abdelrahman Abdallah**[*], **Adam Jatowt**
University of Innsbruck
{abdelrahman.abdallah,adam.jatowt}@uibk.ac.at
laurin.wischounig@student.uibk.ac.at

Abstract

Information Retrieval (IR) is fundamental to many modern NLP applications. The rise of dense retrieval (DR), using neural networks to learn semantic vector representations, has significantly advanced IR performance. Central to training effective dense retrievers through contrastive learning is the selection of informative negative samples. Synthesizing 35 seminal papers, this survey provides a comprehensive and up-to-date overview of negative sampling techniques in dense IR. Our unique contribution is the focus on modern NLP applications and the inclusion of recent Large Language Model (LLM)-driven methods, an area absent in prior reviews. We propose a taxonomy that categorizes techniques including random, static/dynamically mined, and synthetic datasets. We then analyze these approaches with respect to trade-offs between effectiveness, computational cost, and implementation difficulty. The survey concludes by outlining current challenges and promising future directions for the use of LLM-generated synthetic data.

1 Introduction

Information Retrieval (IR) (Baeza-Yates et al., 1999; Bajaj et al., 2016; Gruber et al., 2025; Thakur et al., 2021a) is a foundational field concerned with finding relevant information, typically within large collections of unstructured data (like text documents), that satisfies a user's or system's information need, often expressed as a query. IR powers a multitude of downstream applications including web search, QA and retrieval-augmented generation (RAG).

Traditionally, IR systems relied on lexical sparse retrieval methods, such as Term Frequency-Inverse Document Frequency (TF-IDF) like BM25 (Robertson et al., 2009). These techniques while effective and efficient, often struggle with semantic understanding, vocabulary mismatch (synonyms, paraphrasing), and capturing deeper contextual relationships. In dense retrieval (DR) (Zhao et al., 2024; Chen et al., 2021; Karpukhin et al., 2020b; Sciavolino et al., 2021), queries and documents are encoded into relatively low-dimensional, dense vector representations (embeddings) using neural networks. The relevance score is typically computed based on the similarity (e.g., dot product or cosine similarity) between the query embedding and document embedding in this shared semantic space. Dense retrieval excels at capturing semantic meaning beyond keyword matching. Learning effective dense representations is paramount, and contrastive learning has emerged as a dominant paradigm. This approach trains a model to pull relevant query-document pairs closer together in the embedding space while pushing irrelevant pairs apart.

The strategic selection of these "irrelevant" pairs, referred to as negative sampling (Yang et al., 2024a; Abdallah et al., 2025e,d), is a decisive factor in a model's final performance. The evolution of negative sampling strategies directly maps to performance gains on benchmarks like MS MARCO (Bajaj et al., 2016; Lin et al., 2021a): while random in-batch negatives yield a baseline MRR@10 of 0.261, incorporating static "hard" negatives from BM25 improves it to 0.299. A further leap to 0.330 MRR@10 was achieved by ANCE (Xiong et al., 2020), which introduced dynamic mining to ensure the model is continuously challenged by difficult negatives it finds for itself. This relentless pursuit of harder negatives, however, revealed the false negative problem. The very process of mining top-ranked documents for negatives risks including genuinely relevant but unlabeled passages. RocketQA (Qu et al., 2021) quantified this risk, estimating that 70% of top-retrieved but unlabeled passages are actually positives, a contamination that can poison the training data and severely de-

[*]These authors contributed equally.

Findings of the Association for Computational Linguistics: EACL 2026, pages 3003–3020

grade model performance. This turned out to be such a big problem, that for a long time a lot of research effort on negative sampling in NLP focused on false negative mitigation.

This survey aims to provide a structured and contemporary overview of this critical area. While other reviews exist (Xu et al., 2022; Yang et al., 2024b), they neither focus on negative sampling for dense retrieval nor capture the recent, transformative impact of Large Language Models (LLMs) on negative sampling (Zhao et al., 2025). Our unique contribution is to synthesize the relevant literature with a specific focus on dense retrieval for NLP applications (Li et al., 2024b), by categorizing 35 seminal papers. Also, we focus specifically on negative sampling techniques for contrastive learning in dense retrieval. While we acknowledge complementary approaches like ColBERT's late interaction mechanisms (Khattab and Zaharia, 2020) and knowledge distillation (Lin et al., 2021b), these operate at different levels (architecture vs. sampling strategy) and are appropriately positioned as orthogonal techniques that can be combined with any negative sampling approach discussed here.

The survey addresses three key research questions: **(RQ1)** How can the diverse landscape of negative sampling techniques for dense retrieval be categorized, and what are the core principles, advantages, and trade-offs of each approach? **(RQ2)** How can the diverse landscape of negative sampling techniques for dense retrieval be categorized, and what are the core principles, advantages, and trade-offs of each approach? **(RQ3)** What are the emerging directions and future challenges for negative sampling, particularly with the rise of generative and data-centric methods powered by LLMs?

2 Related Work

General reviews of negative sampling exist for machine learning (Xu et al., 2022) and dense retrieval (Yang et al., 2024b), but neither focuses on dense retrieval methods that have become central to modern NLP applications such as retrieval-augmented generation (RAG), question answering, and dialogue systems. The LLM4IR (Zhu et al., 2023) survey provides broad coverage of LLM applications across IR components, but does not provide a systematic taxonomy focused on negative sampling for contrastive learning, nor does it analyze the critical false negative problem and mitigation strategies that are central to our review. Our survey provides the first comprehensive framework specifically examining negative sampling techniques for dense retrieval in NLP contexts, with emphasis on recent LLM-driven methods absent from prior reviews. We focus exclusively on negative sampling for contrastive learning; complementary approaches like knowledge distillation (Lin et al., 2021b), inference-time query augmentation (Gao et al., 2022a; Abdallah et al., 2025c), late interaction architectures (Khattab and Zaharia, 2020), and refined representations (Ji et al., 2025) are noted but not surveyed in depth. For extended discussion of related work and positioning relative to existing surveys, we refer the reader to Appendix E.

3 Contrastive Learning for Dense Representations

The core idea behind contrastive learning for dense retrieval is to train encoder models E_q (for queries) and E_p (for passages/documents) such that the similarity score $\text{sim}(E_q(q), E_p(p))$ is high for relevant pairs (q, p^+) and low for irrelevant pairs (q, p^-), commonly referred to as triplet loss. Architectures like Sentence-BERT (Reimers and Gurevych, 2019) are commonly trained using such contrastive objectives. A common objective function used is based on Noise Contrastive Estimation (NCE), often implemented as the InfoNCE loss (van den Oord et al., 2019). Given a query q, its positive (relevant) passage p^+, and a set of negative (irrelevant) passages $\{p_i^-\}_{i=1}^N$, let $S = \{p^+\} \cup \{p_i^-\}_{i=1}^N$ be the set of all candidate passages for query q. The goal is to minimize the negative log-likelihood of correctly identifying the positive passage among the candidates:

$$L(q, p^+, S) = \\ -\log \frac{\exp(\text{sim}(E_q(q), E_p(p^+))/\tau)}{\sum_{p \in S} \exp(\text{sim}(E_q(q), E_p(p))/\tau)} \quad (1)$$

Here, $\text{sim}(\cdot, \cdot)$ is a similarity function (e.g., dot product), and τ is a temperature hyperparameter scaling the similarities (Manna et al., 2025). The effectiveness of this learning process hinges on the choice of the negative samples $\{p_i^-\}$. If negatives are too "easy" (semantically very distant from the query), the model receives weak learning signals and may fail to distinguish between relevant passages and challenging irrelevant ones during inference. Conversely, selecting informative "hard" negatives forces the model to learn finer-grained

semantic distinctions, leading to more robust and accurate representations (Zhan et al., 2021a).

4 Taxonomy of Negative Sampling Techniques

Since dense representation learning is a multi-faceted problem, where many aspects of a model can be improved, researchers have proposed many different techniques (Zhan et al., 2020; Lindgren et al., 2021a; Zhan et al., 2021b; Li and Gaussier, 2024). The common objective optimized for are the metrics mentioned above used for quantifying the retrieval performance such as precision, recall and nDCG@k, but also performance on downstream tasks. On the other hand, negative sampling techniques are also often used to improve training efficiency, training stability, sample efficiency, generalizability and domain transferability.

We propose splitting the set of techniques into those that change the way negatives are sampled from the dataset and into ways that are concerned with optimizing the training data itself. The sampling can happen both ahead-of-time as well as during training. Data-centric methods are currently implemented as preprocessing steps, where synthetic data or augmentations are generated before training begins. However, there is no fundamental reason why they could not operate dynamically during training—for instance, generating synthetic hard negatives on-the-fly based on the model's current state, similar to how dynamic mining operates. A high-level overview of the taxonomy is given in Figure 1. Each of these techniques aims to improve one of the above mentioned aspects. Research indicates that most of these techniques can be used in conjunction without sacrificing performance in one aspect to gain another, demonstrating their orthogonality. This has been shown in (Lee et al., 2025), where the authors make use of most of the techniques mentioned in this survey at different stages of their training to create a state-of-the-art embedding model. The unique advantages of the generalized techniques are displayed in Appendix C (Table 10).

4.1 Sampling Techniques

As mentioned before, the goal of negative sampling is to train on samples that provide the best gradients, in order to efficiently train the model. The primary problem of simply training on randomly selected negatives is that it leads to slow convergence, as the model quickly learns to distinguish them. The field has thus evolved a series of sophisticated strategies for selecting more informative negatives. Since the process of mining hard negatives often causes the problem of selecting unlabeled false negatives, the research focus started shifting to the task of avoiding and/or detecting false negatives. We therefore categorize the field of sampling techniques into those focusing on mining hard negatives and those dedicated to mitigating the side effects of this mining.

4.1.1 Random and In-Batch Negatives

The most fundamental strategies leverage readily available passages. Random sampling, where negatives are drawn arbitrarily from the corpus, is simple but inefficient, as it mostly provides "easy" negatives that offer a weak learning signal (Karpukhin et al., 2020a). A more practical and widely adopted approach is In-Batch Negatives (IBNs), which treats all other positive passages within a training mini-batch as negatives for a given query. This method is computationally efficient as it reuses already-processed passages, and has become a standard baseline in many frameworks (Karpukhin et al., 2020a; Reimers and Gurevych, 2019). However, its effectiveness is limited by the batch size, and as training progresses, these negatives often become too easy for the model (Gao et al., 2021; Cheng et al., 2024). Furthermore, IBNs are also susceptible to including accidental false negatives, although this problem is much higher when explicitly mining hard negatives. Despite their disadvantages, In-Batch Negatives are sometimes still used in more sophisticated training regimens as part of the initial phase of training, where the model would not benefit from hard negatives (Lee et al., 2025).

4.1.2 Static Hard Negative Mining

To provide a more consistent challenge, static mining techniques pre-select hard negatives from the corpus in a one-time, offline process before training begins. The most common method uses a sparse retriever like BM25 to find passages that are lexically similar to the query but are not labeled as positive (Karpukhin et al., 2020a). While computationally cheap, this approach can bias the model towards lexical cues and may miss negatives that are semantically challenging but lexically dissimilar. A notable variation, PassageBM25, addresses this by retrieving passages similar to the positive

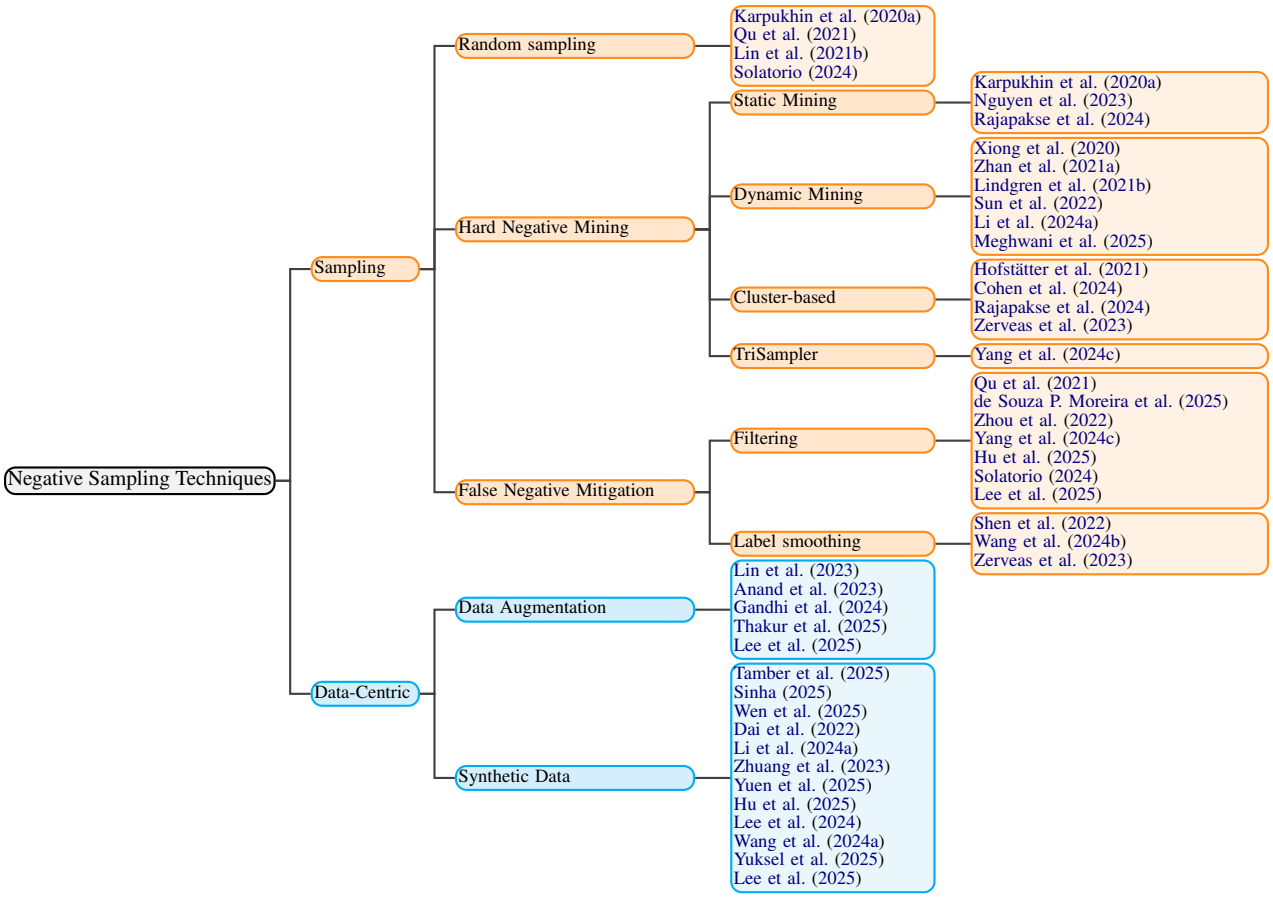

Figure 1: Taxonomy of negative sampling techniques for dense retrieval. The framework divides approaches into two main categories: **Sampling-based techniques** (orange) and **Data-centric techniques** (cyan)

passage p^+ rather than the query, aiming to find documents that are more easily confusable with the correct answer (Nguyen et al., 2023). Without any false negative mitigation techniques, this is very susceptible to unlabeled false negatives.

4.1.3 Dynamic Hard Negative Mining

Dynamic mining represents a significant leap from static methods by using the retrieval model itself to actively find the most challenging negatives throughout the training process. Pioneered by ANCE (Xiong et al., 2020), this model-based technique involves periodically using a recent checkpoint of the dense retriever to re-index the corpus and mine the top-k passages that the current model finds most difficult. This ensures the model is continuously presented with a challenging learning signal, which better aligns the training and inference distributions and leads to superior performance (Zhan et al., 2021a). However, this approach is computationally intensive, requiring repeated indexing and inference, and is highly susceptible to the false negative problem (Qu et al., 2021). The computational load can be slightly reduced by using a cache (Shen et al., 2022). Methods to mitigate the problem of false negatives are discussed in Section 4.2. There are also approaches to improve the qualitative performance of ANCE by trying to predict a point's location in the embedding space in the next step using momentum and lookaheads (Sun et al., 2022). A different idea is to continuously update the negative pool based on the model's performance on concrete samples to maintain a consistent level of difficulty as the model improves (Li et al., 2024a).

4.1.4 Cluster-Based Mining

While model-based dynamic mining excels at finding hard negatives, it can sometimes yield a set of negatives that are semantically very similar to each other. To address this and improve sample diversity, cluster-based mining has been proposed. This approach involves partitioning the document corpus into semantic clusters. Instead of simply taking the top-k hardest negatives, which might all come from the same dense region of the embed-

ding space, negatives are sampled from different clusters. This ensures the model is exposed to a more varied set of challenging examples, forcing it to learn to resolve different types of semantic ambiguity rather than over-specializing on a single type of difficult negative (Cohen et al., 2024; Hofstätter et al., 2021; Zerveas et al., 2023).

4.1.5 Principled Sampling with TriSampler

Moving beyond the simple dichotomy of 'easy' versus 'hard' negatives, TriSampler (Yang et al., 2024c) has focused on establishing more principled criteria for selecting the most informative samples. They introduce the "quasi-triangular principle" to guide negative selection. This principle posits that the ideal negative is not necessarily the absolute hardest one (i.e., the most similar to the query). Instead, it's a sample that is challenging enough to be confusable with the query, but also semantically distinct from the positive passage. Additionally they try to sample negatives with a similar distance to the query as the positive sample to ensure gradients of similar magnitude.

While techniques are categorized by their primary mechanism, understanding their interactions when combined is critical for effective training pipelines. Appendix B analyzes successful combination patterns (e.g., multi-source diversification, knowledge distillation amplification) and quantifies synergistic effects.

4.2 False Negative Mitigation

The pursuit of hard negatives, especially via dynamic mining, creates the false negative problem: mined passages that are actually relevant but unlabeled. Training on these as negatives punishes the model for correct predictions and can severely degrade performance. Consequently, a significant body of research has focused on developing strategies to purify the training signal. These approaches generally fall into three categories: filtering the negative set, regularizing the loss function.

4.2.1 Filtering and Denoising Negatives

The most direct strategy is to filter the mined negative set to remove suspected false negatives. This can be done with simple heuristics, such as top-k filtering, which retains only the most challenging negatives from a larger mined pool (Hu et al., 2025; Zhou et al., 2022; de Souza P. Moreira et al., 2025), or by applying similarity-based thresholds, a technique that can even be used to denoise in-batch negatives (Solatorio, 2024). The aforementioned TriSampler technique (Yang et al., 2024c) sidesteps the need for explicitly removing candidate negatives, since it inherently never samples negatives that are too similar to the query.

Table 1: Impact of false negatives and mitigation effectiveness across datasets.

Dataset	Method	Before FN Mit.	After FN Mit.	Δ
MS MARCO	ANCE→RocketQA	0.330	0.370	+12.1%
TREC-COVID	ANCE	0.654	0.735	+12.4%
Natural Questions	Hard Mining	81.9	84.1	+2.7%

A more powerful, though computationally expensive, solution is denoised hard negative mining. This method, pioneered by RocketQA (Qu et al., 2021), employs a slow but accurate cross-encoder to re-score mined negatives and filter out any identified as likely false negatives. The same idea has been employed in Lee et al. (2025), where the authors used an LLM instead of a cross-encoder to filter out false negatives. This approach clearly is the most powerful, but is associated with enormous compute requirements during training.

4.2.2 Robustness through Regularization

A second class of techniques aims to make the training process itself more robust to noisy labels, regularizing the loss function rather than explicitly altering the data. Contrastive Confidence Regularization, for instance, modifies the loss to prevent the model from becoming overconfident about its negative predictions, thus lessening the penalty from a potential false negative (Wang et al., 2024b). Similarly, the classic technique of label smoothing can be adapted for this purpose. By distributing a small amount of probability mass from the positive sample to the negatives, it reduces the model's sensitivity to any single mislabeled instance. This has proven effective in various contexts, including challenging multilingual settings (Zerveas et al., 2023; Shen et al., 2022; Wang et al., 2024b).

The severity of the false negative problem has been quantified across multiple benchmarks, as shown in Table 1. The results demonstrate that false negative contamination can degrade performance by 10-15% in challenging domains, with mitigation strategies recovering most or all of this loss. The variability in impact across datasets suggests that false negative severity depends on corpus characteristics and query distribution.

4.3 Data-Centric Methods

While the previously discussed negative sampling methods focus on selecting examples from the existing training corpus, data-centric approaches shift the focus to enriching or creating the data itself. These methods operate on the principle that the quality, diversity, and relevance of training data are paramount for model performance. This is achieved either by augmenting existing data or, more extensively, by synthetically generating entirely new datasets.

4.3.1 Data Augmentation

Early data-centric strategies, inspired by techniques in computer vision (Chen et al., 2020), focused on augmenting existing text data. These methods included simple heuristics like replacing keywords with synonyms or mining for additional positive pairs using similarity metrics (Anand et al., 2023). While useful to some degree, these simple techniques have been largely superseded by the more flexible and powerful capabilities of large language models (LLMs).

While not related to negative sampling, there is some research on dynamic query rewriting at inference time to better align queries with the expected structure and content of the positive passages (Baek et al., 2024). The idea of rewriting passages and queries has also been proposed as a form of data augmentation ahead-of-time. This has been explored in (Gandhi et al., 2024) with the explicit goal of tuning datasets to specific tasks. The authors give the example of inverting and rewriting queries and positive passages for code and math tasks. To solve the problem of false negatives (Thakur et al., 2025) propose to create a set of mined negatives and select all the candidate hard negatives using LLMs ahead-of-time.

4.3.2 Synthetic Data for Generalization

LLMs have found various kinds of usage in the training and inference usage of embedding models. When used in the context of dataset augmentation, LLMs mostly find use in a static manner, where they are used to generate synthetic data points before the training starts. Various ways to generate synthetic data using LLMs have been proposed. One way is to only create queries and then create positive matches from the original dataset (Lee et al., 2024). Another way is to forego a base dataset and directly generate the entire dataset synthetically. This has been done in (Wang et al., 2024a; Lee et al., 2025), where they first generate possible tasks and then create positive and hard negative passages for each task.

The advantage of such approaches is that it is possible to tune the generated dataset to the concrete use case of symmetric or asymmetric retrieval. On the other hand, it has been shown that training on various kinds of retrieval scenarios (factoids, opinion questions, short/long queries) improves generalization and overall performance (Tamber et al., 2025; Lee et al., 2025). This claim has also been made by Hu et al. (2025), where the authors generated synthetic samples based on different artificial user personas to increase sample diversity.

Synthetic query generation has also recently been tested in conjunction with other techniques for tasks such as domain adaption. Yuksel et al. (2025) used synthetic data to improve the robustness against domain shifts of dense retrievers when distilling from cross-encoders. Contrary to the points mentioned above, Meghwani et al. (2025) explicitly argues against using synthetic data when training retrieval models for specific and narrow domains, unless data is very limited. Instead they argue for using high-quality datasets and sophisticated hard negative sampling using an ensemble of embedding models and clustering using dimensionality reduction.

5 Evaluation and Empirical Analysis

The development and comparison of negative sampling techniques rely on standardized evaluation protocols, benchmark datasets, and systematic performance analysis. This section establishes the evaluation framework and presents comprehensive empirical results from the surveyed literature.

5.1 Evaluation Metrics

Dense retrieval models are assessed using rank-aware metrics: Mean Reciprocal Rank (MRR) and Normalized Discounted Cumulative Gain (NDCG@k) measure ranking quality, while Recall@k measures the proportion of relevant documents retrieved within top-k results. Metric selection depends on downstream applications—for example, Retrieval-Augmented Generation (RAG) pipelines (Abdallah et al., 2025b,a) prioritize Recall over precision to ensure relevant information reaches the re-ranking or generation stage.

5.2 Benchmark Datasets

Dense retrieval evaluation has evolved from single-task datasets like MS MARCO (Bajaj et al., 2018)

Table 2: MS MARCO Passage Ranking Results showing performance progression across negative sampling techniques.

Method	Technique	Model	MRR@10	R@1k	Source
Random Negatives	In-Batch Random	DPR	0.261	85.4	Karpukhin et al. (2020a)
Static Hard	BM25 Hard Neg.	DPR	0.299	85.4	Karpukhin et al. (2020a)
Dynamic Mining	ANCE	BERT-Siamese	0.330	95.9	Xiong et al. (2020)
Improved Dynamic	TAS-B	BERT	0.347	97.8	Hofstätter et al. (2021)
Advanced Static	TCT-ColBERT	ColBERT	0.359	97.0	Santhanam et al. (2022)
FN Mitigation	RocketQA	ERNIE 2.0	0.370	-	Qu et al. (2021)
Enhanced	RocketQA-v2	ERNIE 2.0	0.388	98.1	Ren et al. (2021)

Table 3: BEIR benchmark zero-shot performance (NDCG@10) across diverse domains. Higher scores indicate better generalization.

Method	Avg. BEIR	TREC-COVID	NFCorpus	Natural Questions	HotpotQA	FiQA
BM25 (Sparse)	42.0	65.6	32.5	32.8	60.3	23.6
DPR (Random)	38.1	65.4	30.1	51.6	54.7	22.8
ANCE (Hard)	39.2	65.4	30.8	52.3	56.8	24.1
SBERT (In-batch)	40.9	59.6	31.3	35.6	57.8	27.9
GTR-Base	44.0	71.6	33.4	36.8	62.2	41.9
Contriever	41.9	59.5	32.8	39.2	56.8	32.3
TAS-B	44.0	-	-	-	-	-

and TREC collections to comprehensive multi-task frameworks. The current standard is MTEB (Massive Text Embedding Benchmark) (Muennighoff et al., 2023), which evaluates embeddings across seven task types: Classification, Clustering, Retrieval, Reranking, Semantic Textual Similarity, Summarization, and Bitext Mining. MTEB incorporates established benchmarks including BEIR (Thakur et al., 2021b)—a heterogeneous collection of IR datasets for assessing zero-shot generalization. This multi-faceted evaluation ensures techniques advance robustly across diverse applications rather than overfitting to single benchmarks.

5.3 Empirical Performance Analysis

To provide systematic insights into the effectiveness of different negative sampling approaches, we present a comprehensive meta-analysis of results reported across the surveyed literature. These comparisons reveal clear performance progression patterns that inform practical deployment decisions.

5.3.1 MS MARCO Passage Ranking

Table 2 summarizes the performance of representative negative sampling techniques on the MS MARCO passage ranking benchmark, demonstrating the clear progression from random to sophisticated sampling strategies. These results demonstrate a clear performance progression: random in-batch negatives establish a baseline (MRR@10: 0.261), static hard negatives provide 14.6% improvement (0.299), dynamic mining yields 26.4% gains (0.330), and false negative mitigation combined with sophisticated mining achieves 41.8% improvement over baseline (0.370+).

Table 4: Natural Questions Open-Domain QA results showing progression from lexical to neural dense retrieval with various negative sampling strategies.

Method	Technique	Top-20	R@100
BM25 Baseline	Lexical	62.9	-
DPR Random	In-Batch	78.4	85.4
ANCE Dynamic	ANN Mining	81.9	87.5
RocketQA	Denoised	84.1	88.5

5.3.2 Natural Questions Open-Domain QA

Table 4 demonstrates technique effectiveness on the Natural Questions benchmark, which tests open-domain question answering capabilities. The progression shows consistent improvement from random negatives (+15.5 accuracy over BM25) to denoised hard negatives (+21.2 over BM25), with false negative mitigation providing an additional 2.2 point gain over unfiltered dynamic mining.

5.3.3 BEIR Zero-Shot Generalization

Table 3 presents zero-shot performance across diverse BEIR datasets, testing model generalization beyond training distribution. The BEIR results reveal that advanced negative sampling techniques (GTR-Base, TAS-B achieving 44.0 average NDCG@10) significantly outperform simpler approaches (DPR at 38.1, ANCE at 39.2). Notably, TAS-B with its combination of knowledge distillation and hard negatives outperforms ANCE on

Table 5: Recent MTEB leaderboard results (2024-2025) showing negative sampling strategies used by top-performing models. Retrieval scores represent average NDCG@10 across MTEB retrieval datasets; Overall scores aggregate performance across 7 task types.

Model	Overall Score	Retrieval Score	Negative Sampling Approach
NV-Embed-v2	69.32	59.36	Hard neg. + Latent attn. + Online mining
Gemini Embed.	66.31	54.36	Multi-stage: Static→Dynamic→LLM denoising→Synthetic
KaLM-Embed.	64.65	51.12	Persona-based synthetic + Ranking filtering
Nomic-Embed	62.28	53.01	Contrastive learning + Hard negatives
BGE-Large-v1.5	60.27	54.29	Hard mining + Cross-batch negatives

14/18 BEIR datasets and DPR on 17/18 datasets, demonstrating the power of technique integration for zero-shot generalization.

5.3.4 Massive Text Embedding Benchmark

Table 5 shows how current state-of-the-art models strategically combine techniques from our taxonomy. Top performers (NV-Embed-v2 at 69.32, Gemini Embeddings at 66.31) achieve results primarily through sophisticated negative sampling strategies rather than architectural innovations alone—all top-5 models employ combinations of static mining, dynamic mining, false negative mitigation, and/or synthetic data generation. The multi-stage progressive approach (Gemini Embeddings: Static→Dynamic→LLM denoising→Synthetic) validates our observation of technique orthogonality, demonstrating that these techniques can be combined synergistically and form the foundation of production systems deployed at scale. Training cost analysis (Table 8, Appendix A) shows static hard negatives provide 6-7% MRR@10 improvement with 33% time increase, while dynamic mining requires 3× cost for 26.4% gains. The multi-stage approach of top models (Static→Dynamic→Denoising→Synthetic) validates technique orthogonality (Appendix B), while ANN index selection guidance for practitioners is provided in Appendix D.

6 Emerging Directions and Future Work

The landscape of negative sampling in Information Retrieval is evolving, with several promising directions emerging from recent research. The increasing sophistication of Large Language Models (LLMs) is undeniably a major driver of these new trends, particularly for synthetic data generation and denoising of mined negatives. In particular, a lot of research leverage the generalization abilities of LLMs to generate diverse and task-specific datasets. Right now the research focuses on the one-off use of LLMs during data set preprocessing/generation. We think that there a lot of unexplored options for dynamic synthetic data generation during training. A concrete example would be to identify subjects or concepts that confuse the model and generate more data points in these domains to add stronger and smoother training signals. The opposite idea would be to dynamically generate adversarial examples in domains, which seem to be well understood by the embedding model. These ideas could be combined with lessons from curriculum learning. Curriculum learning for negative sampling has been explored in (Li et al., 2024a), but not in conjunction with dynamic synthetic data generation. We expect to see more models rely on LLMs for data preprocessing and denoising, as done in (Lee et al., 2025). In theory this does not need to be repeated for each model. Instead, we expect to see more and more synthetic datasets that are ready to use and do not require each research team to synthesize their own dataset. This could in particular lift up the performance of specialized embedding models that focus on domains or languages with limited data.

7 Conclusion

This survey provides the first comprehensive taxonomy of negative sampling techniques for dense retrieval in modern NLP applications, categorizing 35+ papers across random sampling, static/dynamic hard negative mining, false negative mitigation, and LLM-driven synthetic data generation. Our systematic empirical analysis across MS MARCO, Natural Questions, BEIR, and MTEB demonstrates clear performance progression from baseline random negatives (MRR@10: 0.261) to sophisticated combinations achieving 50%+ improvements, with state-of-the-art models (NV-Embed-v2, Gemini Embeddings) validating our framework through strategic technique integration. We quantify critical trade-offs: dynamic mining and denoising require 3-5× training costs but de-

liver proportional performance gains (26-42%), while false negative contamination can degrade performance by 10-15% without mitigation.

8 Limitations

This survey provides a comprehensive overview of negative sampling techniques in dense information retrieval; however, it is important to acknowledge certain limitations.

Firstly, the rapidly evolving nature of research in this domain means that new techniques and refinements are constantly being published. Although we have endeavored to include the most impactful and representative developments to date, it is possible that very recent breakthroughs may not have been fully captured.

Secondly, our focus has primarily been on dense retrieval models, specifically those trained with contrastive learning objectives. Although we briefly touched upon related areas like knowledge distillation and advanced architectures (e.g., ColBERT), a deeper dive into how negative sampling interacts with, or is implicitly handled by, these alternative paradigms was beyond the scope of this survey. For instance, models that rely heavily on generative pre-training or alternative loss functions might employ different strategies for learning discriminative representations that do not directly map to the negative sampling categories discussed here.

Thirdly, while we discussed the importance of various evaluation metrics (Precision, Recall, MAP, NDCG) and common datasets (MS MARCO, TREC collections), a detailed quantitative comparison of all discussed negative sampling techniques across a standardized set of benchmarks was not feasible within the scope of this survey. Such a comparison would require extensive experimental work, which is typically the focus of dedicated benchmarking studies rather than a literature review.

Finally, practical implementation details, and computational trade-offs for each technique were discussed at a high level. The true complexity and performance impact of these methods may vary significantly across different model architectures, datasets, and hardware configurations. Our survey aimed to provide a conceptual understanding and taxonomy rather than a prescriptive guide for implementation.

References

Abdelrahman Abdallah, Mahmoud Abdalla, Bhawna Piryani, Jamshid Mozafari, Mohammed Ali, and Adam Jatowt. 2025a. Rerankarena: A unified platform for evaluating retrieval, reranking and rag with human and llm feedback. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 6593–6597.

Abdelrahman Abdallah, Jamshid Mozafari, Bhawna Piryani, Mohammed Ali, and Adam Jatowt. 2025b. Rankify: A comprehensive python toolkit for retrieval, re-ranking, and retrieval-augmented generation. *arXiv preprint arXiv:2502.02464*.

Abdelrahman Abdallah, Jamshid Mozafari, Bhawna Piryani, and Adam Jatowt. 2025c. Asrank: Zero-shot re-ranking with answer scent for document retrieval. *arXiv preprint arXiv:2501.15245*.

Abdelrahman Abdallah, Jamshid Mozafari, Bhawna Piryani, and Adam Jatowt. 2025d. DeAR: Dual-stage document reranking with reasoning agents via LLM distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5710–5723, Suzhou, China. Association for Computational Linguistics.

Abdelrahman Abdallah, Bhawna Piryani, Jonas Wallat, Avishek Anand, and Adam Jatowt. 2025e. Tempretriever: Fusion-based temporal dense passage retrieval for time-sensitive questions. *arXiv preprint arXiv:2502.21024*.

Abhijit Anand, Jurek Leonhardt, Jaspreet Singh, Koustav Rudra, and Avishek Anand. 2023. Data augmentation for sample efficient and robust document ranking. *Preprint*, arXiv:2311.15426.

Ingeol Baek, Jimin Lee, Joonho Yang, and Hwanhee Lee. 2024. Crafting the path: Robust query rewriting for information retrieval. *Preprint*, arXiv:2407.12529.

R Ricardo Baeza-Yates, Berthier Ribeiro-Neto, and 1 others. 1999. *Modern information retrieval*. ACM Press.

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.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, and 1 others. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *Preprint*, arXiv:2002.05709.

Xilun Chen, Kushal Lakhotia, Barlas Oğuz, Anchit Gupta, Patrick Lewis, Stan Peshterliev, Yashar Mehdad, Sonal Gupta, and Wen-tau Yih. 2021. Salient phrase aware dense retrieval: Can a dense retriever imitate a sparse one? *arXiv preprint arXiv:2110.06918*.

Zesen Cheng, Hang Zhang, Kehan Li, Sicong Leng, Zhiqiang Hu, Fei Wu, Deli Zhao, Xin Li, and Lidong Bing. 2024. Breaking the memory barrier: Near infinite batch size scaling for contrastive loss. *arXiv preprint arXiv:2410.17243*.

Nachshon Cohen, Hedda Cohen Indelman, Yaron Fairstein, and Guy Kushilevitz. 2024. Indi: Informative and diverse sampling for dense retrieval.

Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. *Preprint*, arXiv:2209.11755.

Gabriel de Souza P. Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. 2025. Nv-retriever: Improving text embedding models with effective hard-negative mining. *Preprint*, arXiv:2407.15831.

Saumya Gandhi, Ritu Gala, Vijay Viswanathan, Tongshuang Wu, and Graham Neubig. 2024. Better synthetic data by retrieving and transforming existing datasets. *Preprint*, arXiv:2404.14361.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022a. Precise zero-shot dense retrieval without relevance labels. *Preprint*, arXiv:2212.10496.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022b. Tevatron: An efficient and flexible toolkit for dense retrieval. *Preprint*, arXiv:2203.05765.

Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. 2021. Scaling deep contrastive learning batch size under memory limited setup. *arXiv preprint arXiv:2101.06983*.

Raphael Gruber, Abdelrahman Abdallah, Michael Färber, and Adam Jatowt. 2025. ComplexTempQA: A 100m dataset for complex temporal question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9100–9112, Suzhou, China. Association for Computational Linguistics.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. *Preprint*, arXiv:2104.06967.

Xinshuo Hu, Zifei Shan, Xinping Zhao, Zetian Sun, Zhenyu Liu, Dongfang Li, Shaolin Ye, Xinyuan Wei, Qian Chen, Baotian Hu, Haofen Wang, Jun Yu, and Min Zhang. 2025. Kalm-embedding: Superior training data brings a stronger embedding model. *Preprint*, arXiv:2501.01028.

Yifan Ji, Zhipeng Xu, Zhenghao Liu, Yukun Yan, Shi Yu, Yishan Li, Zhiyuan Liu, Yu Gu, Ge Yu, and Maosong Sun. 2025. Learning more effective representations for dense retrieval through deliberate thinking before search. *Preprint*, arXiv:2502.12974.

Vladimir Karpukhin, Barlas Oguz, Sewon Chen, Patrick Lischinski, Armand Joulin, and Edouard Grave. 2020a. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020b. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, Xiaoqi Ren, Shanfeng Zhang, Daniel Salz, Michael Boratko, Jay Han, Blair Chen, Shuo Huang, Vikram Rao, Paul Suganthan, and 28 others. 2025. Gemini embedding: Generalizable embeddings from gemini. *Preprint*, arXiv:2503.07891.

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernandez Abrego, Weiqiang Shi, Nithi Gupta, Aditya Kusupati, Prateek Jain, Siddhartha Reddy Jonnalagadda, Ming-Wei Chang, and Iftekhar Naim. 2024. Gecko: Versatile text embeddings distilled from large language models. *Preprint*, arXiv:2403.20327.

Minghan Li and Eric Gaussier. 2024. Domain adaptation for dense retrieval and conversational dense retrieval through self-supervision by meticulous pseudo-relevance labeling. *arXiv preprint arXiv:2403.08970*.

Shiyu Li, Yang Tang, Shizhe Chen, and Xi Chen. 2024a. Conan-embedding: General text embedding with more and better negative samples. *Preprint*, arXiv:2408.15710.

Xiaopeng Li, Xiangyang Li, Hao Zhang, Zhaocheng Du, Pengyue Jia, Yichao Wang, Xiangyu Zhao, Huifeng

Guo, and Ruiming Tang. 2024b. Syneg: Llm-driven synthetic hard-negatives for dense retrieval. *Preprint*, arXiv:2412.17250.

Sheng-Chieh Lin, Jheng-Hong Yang, Min-Yen Chen, and Jimmy Lin. 2023. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. *arXiv preprint arXiv:2302.07452*.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021a. In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 163–173.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021b. In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*.

Erik Lindgren, Sashank Reddi, Ruiqi Guo, and Sanjiv Kumar. 2021a. Efficient training of retrieval models using negative cache. *Advances in Neural Information Processing Systems*, 34:4134–4146.

Erik Lindgren, Sashank Reddi, Ruiqi Guo, and Sanjiv Kumar. 2021b. Efficient training of retrieval models using negative cache. In *Advances in Neural Information Processing Systems*, volume 34, pages 4134–4146. Curran Associates, Inc.

Siladittya Manna, Soumitri Chattopadhyay, Rakesh Dey, Umapada Pal, and Saumik Bhattacharya. 2025. Dynamically scaled temperature in self-supervised contrastive learning. *IEEE Transactions on Artificial Intelligence*.

Hansa Meghwani, Amit Agarwal, Priyaranjan Pattnayak, Hitesh Laxmichand Patel, and Srikant Panda. 2025. Hard negative mining for domain-specific retrieval in enterprise systems. *Preprint*, arXiv:2505.18366.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. *Preprint*, arXiv:2210.07316.

Thanh-Do Nguyen, Chi Minh Bui, and Xuan-Hieu Phan. 2023. Passage-based bm25 hard negatives: A simple and effective negative sampling strategy for dense retrieval. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation (PACLIC 2023)*.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. *Preprint*, arXiv:2010.08191.

Thilina Chaturanga Rajapakse, Andrew Yates, and Maarten de Rijke. 2024. Negative sampling techniques for dense passage retrieval in a multilingual setting. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 575–584, New York, NY, USA. Association for Computing Machinery.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. Rocketqav2: A joint training method for dense passage retrieval and passage re-ranking. *arXiv preprint arXiv:2110.07367*.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Keshav Santhanam, Omar Khattab, Christopher Potts, and Matei Zaharia. 2022. Plaid: an efficient engine for late interaction retrieval. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1747–1756.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. *arXiv preprint arXiv:2109.08535*.

Tianhao Shen, Mingtong Liu, Ming Zhou, and Deyi Xiong. 2022. Recovering gold from black sand: Multilingual dense passage retrieval with hard and false negative samples. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10659–10670, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aarush Sinha. 2025. Don't retrieve, generate: Prompting llms for synthetic training data in dense retrieval. *Preprint*, arXiv:2504.21015.

Aivin V. Solatorio. 2024. Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning. *Preprint*, arXiv:2402.16829.

Si Sun, Chenyan Xiong, Yue Yu, Arnold Overwijk, Zhiyuan Liu, and Jie Bao. 2022. Reduce catastrophic forgetting of dense retrieval training with teleportation negatives. *Preprint*, arXiv:2210.17167.

Manveer Singh Tamber, Suleman Kazi, Vivek Sourabh, and Jimmy Lin. 2025. Conventional contrastive learning often falls short: Improving dense retrieval with cross-encoder listwise distillation and synthetic data. *Preprint*, arXiv:2505.19274.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021a. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021b. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *Preprint*, arXiv:2104.08663.

Nandan Thakur, Crystina Zhang, Xueguang Ma, and Jimmy Lin. 2025. Fixing data that hurts performance: Cascading llms to relabel hard negatives for robust information retrieval. *Preprint*, arXiv:2505.16967.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving text embeddings with large language models. *Preprint*, arXiv:2401.00368.

Shiqi Wang, Yeqin Zhang, and Cam-Tu Nguyen. 2024b. Mitigating the impact of false negatives in dense retrieval with contrastive confidence regularization. *Preprint*, arXiv:2401.00165.

Haoyang Wen, Jiang Guo, Yi Zhang, Jiarong Jiang, and Zhiguo Wang. 2025. On synthetic data strategies for domain-specific generative retrieval. *Preprint*, arXiv:2502.17957.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *Preprint*, arXiv:2007.00808.

Lanling Xu, Jianhai Luo, and Jianhua Li. 2022. Negative sampling for contrastive representation learning: A review. *arXiv preprint arXiv:2206.00212*.

Zhen Yang, Ming Ding, Tinglin Huang, Yukuo Cen, Junshuai Song, Bin Xu, Yuxiao Dong, and Jie Tang. 2024a. Does negative sampling matter? a review with insights into its theory and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5692–5711.

Zhen Yang, Haochen Ma, Xin Sun, Jian-Yun Yu, Jiajie Lin, Hongtao Chen, Zicong Song, and Ji-Rong Wen. 2024b. Does negative sampling matter? a review with insights into its theory and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhen Yang, Zhou Shao, Yuxiao Dong, and Jie Tang. 2024c. Trisampler: A better negative sampling principle for dense retrieval. *Preprint*, arXiv:2402.11855.

Sizhe Yuen, Ting Su, Ziyang Wang, Yali Du, and Adam J. Sobey. 2025. Automatic dataset generation for knowledge intensive question answering tasks. *Preprint*, arXiv:2505.14212.

Goksenin Yuksel, David Rau, and Jaap Kamps. 2025. Remining hard negatives for generative pseudo labeled domain adaptation. *Preprint*, arXiv:2501.14434.

George Zerveas, Navid Rekabsaz, and Carsten Eickhoff. 2023. Enhancing the ranking context of dense retrieval methods through reciprocal nearest neighbors. *Preprint*, arXiv:2305.15720.

Jingtao Zhan, Jiaxin Dai, Yizhen Ding, Yujie Huang, Weijiang Liu, Shujian Chu, Yi Wang, Daxin Huang, and Yunjie Zhou. 2021a. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021b. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1503–1512.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. Learning to retrieve: How to train a dense retrieval model effectively and efficiently. *arXiv preprint arXiv:2010.10469*.

Chu Zhao, Enneng Yang, Yuting Liu, Jianzhe Zhao, Guibing Guo, and Xingwei Wang. 2025. Can llm-driven hard negative sampling empower collaborative filtering? findings and potentials. *arXiv preprint arXiv:2504.04726*.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4):1–60.

Kun Zhou, Yu Fu, Guangwei Luo, Huazheng Zhu, Yifan Chen, Xiangbo Shi, Hongwei Zhou, Yang Zhang, and Jun Guan. 2022. Simans: Simple ambiguous negatives sampling for dense text retrieval. *arXiv preprint arXiv:2210.11773*.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *CoRR*, abs/2308.07107.

Shengyao Zhuang, Houxing Ren, Linjun Shou, Jian Pei, Ming Gong, Guido Zuccon, and Daxin Jiang. 2023. Bridging the gap between indexing and retrieval for differentiable search index with query generation. *Preprint*, arXiv:2206.10128.

A Computational Cost-Benefit Analysis

Understanding the trade-offs between performance gains and computational requirements is critical for practical deployment. We present unified analysis using standardized protocols and real-world measurements.

A.1 Training Cost Analysis

Table 8 presents wall-clock training times from the Tevatron toolkit (Gao et al., 2022b) on matched hardware (4×A100 GPUs and TPU v3-8) using MS MARCO Passage dev set, providing fair comparison across techniques. These results demonstrate that adding static hard negatives provides 6-7% absolute MRR@10 improvement (roberta-large: 0.339→0.361) with approximately 33% training time increase. The co-condenser backbone with hard negatives achieves the best efficiency, reaching 0.382 MRR@10 in only 4 GPU hours. Dynamic mining methods like ANCE achieve 0.330 MRR@10 with significantly higher computational requirements (3x baseline) due to periodic re-indexing. Table 6 summarizes the cost-benefit trade-offs across major technique categories. This analysis reveals that while sophisticated techniques like dynamic mining and denoising require 3-5x training costs, they deliver proportionally larger performance gains (26-42%), making them worthwhile for production systems where retrieval quality is critical.

Table 6: Computational cost-benefit analysis of negative sampling techniques. Training time measured relative to random in-batch baseline.

Technique	Training Time Multiplier	Memory Overhead	Impl. Difficulty	Perf. Gain (MRR@10)
Random In-Batch	1.0x	Minimal	Easy	Baseline
Static BM25	1.1x	Low	Medium	+14.6%
Dynamic ANCE	3.0x	High	Hard	+26.4%
Denoised (RocketQA)	4.5x	Very High	Medium	+41.8%
LLM Synthetic	2.0x (one-time)	Medium	Hard	Variable
Combined Approach	5.0x+	Very High	Very Hard	+50%+

A.2 Inference Scalability and Latency

Production deployment requires understanding retrieval latency at scale. We analyze two architectural paradigms: late interaction models (multi-vector) and dense single-vector retrievers.

Late Interaction Systems. ColBERTv2 with the PLAID engine (Santhanam et al., 2022) demonstrates that multi-vector scoring can achieve production-scale performance. Table 7 shows measured latency on 140M passage corpus.

Table 7: ColBERTv2 with PLAID engine: latency and speedup at 140M passages.

Engine	Hardware	Latency	Speedup
ColBERTv2 baseline	GPU	250 ms	1×
PLAID	GPU	35-100 ms	2.5-7×
PLAID	CPU	150-300 ms	9-45×

Table 8: Unified empirical comparison of effectiveness vs. training cost. All results from Tevatron toolkit on matched setup (4×A100 / TPU v3-8) with identical preprocessing, batch size, and epochs.

Model / Training	Negatives	MRR@10	GPU time	TPU time
distilbert-base	in-batch	0.316	1.5 h	1.0 h
bert-base	in-batch	0.322	3.0 h	2.0 h
co-condenser-marco	in-batch	0.357	3.0 h	2.0 h
bert-large	in-batch	0.327	7.5 h	6.0 h
roberta-large	in-batch	0.339	7.5 h	6.0 h
roberta-large	static HN	0.361	10.0 h	8.0 h
co-condenser-marco	static HN	**0.382**	**4.0 h**	**3.0 h**

PLAID achieves 2.5-7× GPU speedup and 9-45× CPU speedup over vanilla ColBERTv2, reaching tens of milliseconds on GPU and hundreds of milliseconds on CPU at 140M passages without quality loss. This demonstrates that late interaction models, which reduce reliance on hard negative mining during training by deferring scoring to inference time, are viable for production deployment.

Single-Vector Dense Retrieval Indexing. For traditional dense retrievers, approximate nearest neighbor (ANN) index choice critically impacts latency-quality trade-offs.

B Technique Interaction Analysis

While the taxonomy in Section C categorizes techniques by their primary mechanism, understanding how these techniques interact when combined is critical for building effective training pipelines. Analysis of successful implementations reveals clear patterns of synergistic combinations.

B.1 Successful Combination Patterns

Recent state-of-the-art models demonstrate that techniques from different categories can be effectively integrated without conflict. We identify three primary interaction patterns:

Pattern 1: Multi-Source Negative Diversification Qu et al. (2021) demonstrates effective integration by combining multiple negative sources: cross-batch negatives to increase pool size, denoised hard negatives to remove false positives, and data augmentation through synthetic queries. This addresses different training challenges simultaneously—cross-batch increases diversity, denoising ensures quality, and synthesis adds domain coverage. The result is a 12.1% improvement over ANCE (MRR@10: 0.370 vs 0.330), showing that these approaches complement rather than interfere with each other.

Pattern 2: Knowledge Distillation Amplification Hofstätter et al. (2021) combines in-batch negatives with knowledge distillation using Margin-MSE loss and BM25 hard negatives. The key insight is that knowledge distillation provides better training signals while hard negatives increase difficulty level. This synergy results in superior zero-shot generalization on BEIR (NDCG@10: 44.0 vs 39.2 for ANCE), demonstrating that distillation amplifies the effectiveness of negative sampling rather than replacing it.

Pattern 3: Progressive Difficulty Scaling Lee et al. (2025) implements a multi-stage approach: Static BM25 $\rightarrow$ Dynamic ANCE-style mining $\rightarrow$ LLM denoising $\rightarrow$ Synthetic data augmentation. This progressive pipeline allows the model to first learn basic distinctions (static negatives), then adapt to harder examples (dynamic mining), followed by quality refinement (denoising), and finally domain expansion (synthetic data). The staged approach prevents early training instability while maximizing final performance.

B.2 Key Interaction Principles

Analysis of these successful combinations reveals several general principles: (1) **False negative mitigation enables aggressive mining**: Denoising techniques make dynamic hard negative mining safe by removing contamination. Without mitigation, aggressive mining degrades performance; with mitigation, it provides the strongest learning signal. (2) **Multi-source negatives reduce overfitting**: Combining different negative sources (in-batch, BM25, dynamic, synthetic) prevents the model from overfitting to any single retrieval pattern, improving domain transfer. (3) **Synthetic data complements rather than replaces mining**: LLM-generated synthetic queries/passages work best when combined with mined hard negatives from real data, providing both breadth (synthetic) and specificity (mined). (4) **Curriculum matters for complex combinations**: When using multiple techniques, starting with simpler methods (in-batch, static) and progressively adding sophisticated approaches (dynamic, synthetic) improves training stability.

B.3 Anti-patterns to Avoid

Not all combinations are beneficial. Our analysis identifies potential conflicts: (1) **Dynamic mining without denoising**: The false negative contamination in dynamic mining (up to 70% according to RocketQA) can negate its benefits if not filtered. (2) **Excessive negative pool sizes**: Combining multiple hard negative sources without careful sampling can create excessively large negative sets that dilute the learning signal and increase computational cost without proportional gains. (3) **Premature hard negative introduction**: Starting training immediately with very hard negatives can cause instability. Models benefit from an initial warm-up phase with easier negatives.

Table 9: Effectiveness of technique combinations. Performance improvements shown relative to single-technique baselines.

Technique Combination	MS MARCO MRR@10	BEIR Avg. NDCG@10
In-batch only (baseline)	0.322	40.9
+ Static BM25	0.361 (+12.1%)	41.5 (+1.5%)
+ Dynamic mining	0.330 (+2.5%)	39.2 (-4.2%)
+ Dynamic + Denoising	0.370 (+14.9%)	42.8 (+4.6%)
+ Multi-source + Distillation	0.347 (+7.8%)	44.0 (+7.6%)
+ Progressive (Gemini)	0.382 (+18.6%)	45.2 (+10.5%)

Table 9 quantifies the impact of technique combinations across benchmarks. This analysis provides practitioners with evidence-based guidance: complementary approaches following established patterns (diversification, amplification, progression) consistently achieve superior performance, while naive combination without consideration of interactions can harm results.

C Negative Sampling Techniques

This appendix provides a consolidated quick-reference table summarizing all negative sampling techniques discussed in this survey. While the full taxonomy (Figure 1) shows the hierarchical organization and associated papers, Table 10 focuses on practical selection criteria: each technique's unique advantage, computational cost, and implementation difficulty.

Computational Cost Ratings:

- **Very Low**: Negligible overhead beyond baseline training (e.g., in-batch negatives)
- **Low**: One-time preprocessing cost, minimal runtime overhead (e.g., static BM25 mining)
- **Medium**: Moderate overhead from clustering or synthetic generation (1.5-2× baseline)
- **High**: Significant overhead from repeated re-indexing or cross-encoder filtering (3-5× baseline)

Table 10: Overview of Negative Sampling Techniques in Dense Information Retrieval. This table details each technique's primary advantage, re-evaluated computational cost, and implementation difficulty. Refer to the full taxonomy for a complete list of surveyed papers.

Type	Technique	Unique Advantage	Comp. Cost	Impl. Difficulty
Sampling-Based Techniques				
Random Sampling	Random In-Batch Negatives	Computationally free as it reuses passages within a batch. Effective with large batch sizes (Karpukhin et al., 2020a; Abdallah et al., 2025e; Reimers and Gurevych, 2019).	Very Low	Easy
Hard Negative Mining	Static	Finds lexically similar hard negatives with a one-time, offline cost. Can be based on query similarity (BM25) or positive passage similarity (PassageBM25) (Karpukhin et al., 2020a; Nguyen et al., 2023).	Low	Medium
Hard Negative Mining	Dynamic Model-Based	Continuously finds the hardest negatives according to the current model state, ensuring a challenging and relevant learning signal throughout training (Xiong et al., 2020).	High	Hard
Hard Negative Mining	Cluster-Based	Samples negatives from different semantic clusters to ensure diversity and avoid redundant hard negatives, focusing on semantic ambiguity (Cohen et al., 2024; Zerveas et al., 2023).	Medium	Hard
Hard Negative Mining	TriSampler	Sample negatives from the quasi-triangular region created by the query and the positive sample to ensure uniform gradients (Yang et al., 2024c).	Medium	Medium
False Negative Mitigation	Top-k Filtering	Assume the top-k most relevant potential negatives are false negatives and ignore them when sampling negatives (Zhou et al., 2022; Hu et al., 2025; de Souza P. Moreira et al., 2025; Solatorio, 2024).	Very Low	Easy
False Negative Mitigation	Denoised Hard Negatives	Explicitly filters false negatives from the mined set, using a powerful but slow cross-encoder or LLM, to purify the training signal (Qu et al., 2021; Lee et al., 2025).	High	Medium
False Negative Mitigation	Label Smoothing/Contrastive Confidence Regularization	Mitigates the impact of false negatives via loss regularization, improving robustness without requiring an explicit and costly filtering step (Wang et al., 2024b).	Low	Medium
Data-Centric Techniques (LLM-based)				
Data-Centric	Data Augmentation	Extend existing datasets by replacing keywords in passages with synonyms (Anand et al., 2023).	Medium (one-time)	Medium
Data-Centric	Synthetic Data	Generates synthetic queries, passages and/or labels using LLMs for diverse and specific retrieval tasks (e.g., symmetric/asymmetric, domain-specific, persona-based) to improve model generalization and robustness (Tamber et al., 2025; Hu et al., 2025).	High (one-time)	Hard

Implementation Difficulty Ratings:

- **Easy**: Can be implemented with standard libraries in <100 lines of code
- **Medium**: Requires moderate engineering (e.g., BM25 integration, filtering pipelines)
- **Hard**: Requires substantial infrastructure (e.g., dynamic re-indexing, LLM pipelines, clustering)

Practitioners should use this table to identify techniques matching their computational budget and engineering capacity. As demonstrated in Section , techniques from different categories can be combined synergistically—for example, static mining (Low cost, Medium difficulty) with top-k filtering (Very Low cost, Easy) provides substantial gains without requiring advanced infrastructure.

D Index Selection Guide for Dense Retrieval Systems

When deploying dense retrieval systems at scale, the choice of Approximate Nearest Neighbor (ANN) index is critical for balancing retrieval quality, latency, and memory consumption. This appendix provides detailed guidance on index selection with parameter definitions to help practitioners configure their systems effectively.

Table 11 summarizes the trade-offs between major indexing approaches. The choice depends on corpus size, available memory, latency requirements, and acceptable recall levels. For corpora under 10M vectors with sufficient memory, exact search (Flat index) is preferred. For larger scales, approximate methods become necessary.

D.1 Parameter Definitions and Tuning Guidelines

IVF-Flat (Inverted File with Flat Quantization)

- **`nlist`**: Number of Voronoi cells (clusters) to partition the vector space. Typical range: 100–10,000. Higher values reduce search space per query but increase quantization error. Rule of thumb: $\sqrt{N}$ where N is corpus size, or $4\sqrt{N}$ for better recall.
- **`nprobe`**: Number of cells to visit during search. Range: 1–100+. Higher values improve recall at cost of latency. Start with `nprobe=1` and increase until desired recall achieved. Typical production: 8–32.

IVF-PQ (Inverted File with Product Quantization)

- **`nlist`**, **`nprobe`**: Same as IVF-Flat above.
- **`m`**: Number of subquantizers (segments to split each vector into). Must divide vector dimension evenly. Typical: 8, 16, 32, 64. Higher m preserves more information but increases memory. For 768-dim vectors, $m = 8$ gives 96-dim per segment.
- **`nbits`**: Bits per subquantizer code. Typical: 8 bits (256 centroids per subquantizer). Determines codebook size: 2^{nbits} centroids. Memory per vector: $m \times$ nbits bits.

HNSW (Hierarchical Navigable Small World)

- **`M`**: Number of bi-directional links created for each node. Range: 16–64. Higher values improve recall and reduce hops but increase memory ($\sim M \times 2 \times 4$ bytes per vector for pointers). Typical: 16–32 for production.
- **`efConstruction`**: Size of dynamic candidate list during index construction. Range: 100–500. Higher values create better graphs but slower indexing. Should be $\geq$ `efSearch`. Typical: 200–400.
- **`efSearch`**: Size of dynamic candidate list during query. Range: 50–500. Directly controls recall-latency trade-off. Start with 100 and tune. Does not affect index size, only search time. Typical production: 64–128.

ScaNN (Scalable Nearest Neighbors - Google Research)

- **`num_leaves`**: Number of leaf partitions in the tree. Similar to `nlist` in IVF. Typical: 1000–10,000. Determines granularity of vector space partitioning.
- **`anisotropic_quantization_threshold`**: Threshold for applying anisotropic vector quantization, which adapts quantization to data distribution. Range: 0.0–1.0. Lower values apply quantization more aggressively.
- **`dimensions_per_block`**: For product quantization, number of dimensions per quantization block. Typical: 2–8.

SOAR (Spilling with Orthogonality-Amplified Residuals)

- **`redundancy_factor`**: Controls how many times each vector is replicated across partitions. Range: 1.5–5.0. Higher redundancy improves recall by allowing vectors to be found via multiple paths, at cost of index size. Typical: 2.0–3.0 for 2–3× speedup over ScaNN baseline.
- **`replication_count`**: Explicit number of partition replications per vector. Alternative to `redundancy_factor`. Typical: 2–4 partitions.

D.2 Selection Decision Tree

1. **Corpus size < 10M vectors**: Use Flat index for exact search if memory permits.
2. **Memory constrained (billion-scale corpora)**: Use IVF-PQ with aggressive compression ($m = 8$, nbits $= 8$).
3. **Recall priority with adequate memory**: Use HNSW with $M = 32$, tune `efSearch` for recall target.
4. **Latency priority with budget**: Use ScaNN or SOAR with appropriate redundancy.
5. **Trillion-scale (Google/Meta scale)**: Use IVF-PQ with distributed sharding, as reported by Faiss team.

Table 11: ANN index comparison for single-vector dense retrieval. Trade-offs between memory, latency, and recall at scale.

Index Type	Memory	Latency	Recall	Key Parameters
Flat (Exact)	Very High	Low (10M scale)	100%	None (exhaustive search)
IVF-Flat	Medium	Medium	95-99%	`nlist`, `nprobe`
IVF-PQ	Low	Medium-High	90-95%	`nlist`, `nprobe`, `m`, `nbits`
HNSW	High	Very Low	95-99%	`M`, `efConstruction`, `efSearch`
ScaNN	Medium-High	Low	95-99%	`num_leaves`, `anisotropic_quantization_threshold`
SOAR (ScaNN+)	Medium	Very Low	95-99%	`redundancy_factor`, `replication_count`

E Related Work and Survey

This appendix provides detailed positioning of our survey relative to existing literature and explains our scope decisions.

E.1 Comparison with Existing Surveys

General Negative Sampling Surveys. Xu et al. (2022) provides a comprehensive review of negative sampling across machine learning, covering applications in recommendation systems, graph neural networks, and metric learning. However, their treatment of information retrieval is limited and predates the LLM era. Similarly, Yang et al. (2024b) surveys negative sampling for dense retrieval but focuses primarily on computer vision and cross-modal retrieval, with minimal coverage of text-only dense retrieval or modern NLP applications.

LLM4IR Survey. The LLM4IR (Zhu et al., 2023) survey offers broad coverage of how LLMs can be applied across all IR components: query rewriting, retrieval, reranking, and reading comprehension. While valuable, it treats LLM applications holistically rather than providing systematic analysis of any single training component. Specifically, LLM4IR mentions data augmentation techniques but does not: (1) categorize these through a negative sampling framework, (2) analyze the false negative problem and mitigation strategies, (3) examine computational cost vs. effectiveness trade-offs, (4) provide systematic performance comparisons, or (5) discuss technique combinations and interactions. Our survey complements LLM4IR by providing technical depth on the contrastive learning optimization process.

Multilingual IR Study. Rajapakse et al. (2024) compares multiple negative sampling methods specifically for multilingual IR, examining in-distribution and out-of-distribution performance. While this empirical study is valuable, it is not a comprehensive survey and focuses narrowly on cross-lingual transfer rather than the broader taxonomy of techniques.

E.2 Complementary Approaches Not Surveyed

Our survey focuses exclusively on negative sampling for contrastive learning in dense retrieval. Several complementary approaches improve retrieval quality but operate at different levels:

Knowledge Distillation. Training bi-encoders to mimic more powerful cross-encoder scores provides quality improvements orthogonal to negative sampling. The interaction between in-batch negatives and knowledge distillation has been studied (Lin et al., 2021b), but distillation itself is a teacher-student paradigm rather than a data selection strategy.

Inference-Time Query Augmentation. Methods like HyDE (Gao et al., 2022a) use LLMs to generate hypothetical documents from queries at inference time, improving retrieval without modifying training. Similarly, Lin et al. (2023) uses first-pass retrieval to rewrite queries for second-pass retrieval. These approaches enhance query representation rather than training data selection.

Alternative Architectures. ColBERT (Khattab and Zaharia, 2020) uses late interaction mechanisms, deferring fine-grained scoring to inference time. This shifts complexity away from training-time negative sampling, but ColBERT still requires contrastive training with negatives. Such architectural innovations are complementary to but distinct from negative sampling strategies.

Refined Representations. Methods like DEBATER (Ji et al., 2025) use LLM reasoning to create nuanced document representations before retrieval. These preprocessing approaches can be combined with any negative sampling technique.

E.3 Scope Justification

Our focused scope on negative sampling for contrastive learning serves researchers and practitioners who need to optimize this critical training component. Deep expertise in negative sampling optimization—covering random sampling, static/dynamic mining, false negative mitigation, and LLM-based synthesis—provides more value to this audience than shallow coverage of all retrieval approaches. Recent SOTA models (NV-Embed-v2, Gemini Embeddings) achieve their performance primarily through negative sampling innovations rather than architectural changes, validating the importance of our focus area.