

# Hard Negative Mining for Domain-Specific Retrieval in Enterprise Systems

Hansa Meghwani\*, Amit Agarwal\*, Priyaranjan Pattnayak,  
Hitesh Laxmichand Patel, Srikant Panda

Oracle AI

**Correspondence:** [hansa.meghwani@oracle.com](mailto:hansa.meghwani@oracle.com); [amit.h.agarwal@oracle.com](mailto:amit.h.agarwal@oracle.com) \*

## Abstract

Enterprise search systems often struggle to retrieve accurate, domain-specific information due to semantic mismatches and overlapping terminologies. These issues can degrade the performance of downstream applications such as knowledge management, customer support, and retrieval-augmented generation agents. To address this challenge, we propose a scalable hard-negative mining framework tailored specifically for domain-specific enterprise data. Our approach dynamically selects semantically challenging but contextually irrelevant documents to enhance deployed re-ranking models.

Our method integrates diverse embedding models, performs dimensionality reduction, and uniquely selects hard negatives, ensuring computational efficiency and semantic precision. Evaluation on our proprietary enterprise corpus (cloud services domain) demonstrates substantial improvements of 15% in MRR@3 and 19% in MRR@10 compared to state-of-the-art baselines and other negative sampling techniques. Further validation on public domain-specific datasets (FiQA, Climate Fever, TechQA) confirms our method’s generalizability and readiness for real-world applications.

## 1 Introduction

Accurate retrieval of domain-specific information significantly impacts critical enterprise processes, such as knowledge management, customer support, and Retrieval Augmented Generation (RAG) Agents. However, achieving precise retrieval remains challenging due to semantic mismatches, overlapping terminologies, and ambiguous abbreviations common in specialized fields like finance, and cloud computing. Traditional lexical retrieval techniques, such as BM25 (Robertson and Walker, 1994), struggle due to vocabulary mismatches, leading to irrelevant results and poor user experience.

Recent dense retrieval approaches leveraging pre-trained language models, like BERT-based encoders (Karpukhin et al., 2020; Xiong et al., 2020; Guu et al., 2020), mitigate lexical limitations by capturing semantic relevance. Nevertheless, their performance heavily relies on the negative samples—documents incorrectly retrieved due to semantic similarity but lacking contextual relevance. Models trained with negative sampling methods (e.g., random sampling, BM25-based static sampling, or dynamic methods like ANCE (Xiong et al., 2020), STAR (Zhan et al., 2021)) either lack sufficient semantic discrimination or incur high computational costs, thus limiting scalability and practical enterprise deployment. For instance, given a query such as “*Steps to deploy a MySQL database on Cloud Infrastructure*,” most negative sampling techniques select documents discussing non-MySQL database deployments. Conversely, our method strategically selects a hard negative discussing MySQL deployment on-premises, which despite semantic overlap, is contextually distinct and thus poses a stronger training challenge for the retrieval and re-ranking models.

Our proposed framework addresses these by introducing a novel semantic selection criterion explicitly designed to curate high-quality hard negatives. By uniquely formulating two semantic conditions that effectively select negatives that closely resemble query semantics but remain contextually irrelevant, significantly minimizing false negatives encountered by existing techniques. The main contributions of this paper are:

1. A negative mining framework for dynamically selecting semantically challenging hard negatives, leveraging diverse embedding models and semantic filtering criteria to significantly improve re-ranking models in domain-specific retrieval scenarios.
2. Comprehensive evaluations demonstrating

\*The authors contributed equally to this work.

consistent and significant improvements across both proprietary and publicly available datasets, verifying our method’s impact and broad applicability across domain-specific usecases.

3. In-depth analysis, of critical challenges in handling both short and long-form enterprise documents, laying a clear foundation for targeted future improvements.

Our work directly enhances the semantic discrimination capabilities of re-ranking models, resulting in **15% improvement in MRR@3** and **19% improvement in MRR@10** on our in-house cloud-services domain dataset. Further evaluations on public domain-specific benchmarks (FiQA, Climate Fever, TechQA) confirm generalizability and tangible improvements of our proposed negative mining framework.

## 2 Related Work

### 2.1 Hard Negatives in Retrieval Models

The role of hard negatives in training dense retrieval models has been widely studied. Static negatives, such as BM25 (Robertson and Walker, 1994), provide lexical similarity but fail to capture semantic relevance, often leading to overfitting (Qu et al., 2020). Dynamic negatives, introduced in ANCE (Xiong et al., 2020) and STAR (Zhan et al., 2021), adapt during training to provide more challenging contrasts but require significant computational resources due to periodic re-indexing. Our framework addresses these limitations by dynamically identifying semantically challenging negatives using clustering and dimensionality reduction, ensuring scalability and adaptability.

Further studies have explored advanced methods for negative sampling in cross-encoder models (Meghwani, 2024). Localized Contrastive Estimation (LCE) (Guo et al., 2023) integrates hard negatives into cross-encoder training, improving the reranking performance when negatives align with the output of the retriever. Similarly, (Pradeep et al., 2022) demonstrated the importance of hard negatives even when models undergo advanced pre-training techniques, such as condenser (Gao and Callan, 2021). Our work builds on these efforts by offering a scalable approach, which can be applied to any domain-heavy enterprise data.

### 2.2 Negative Sampling Strategies

Effective negative sampling significantly affects the performance of the retrieval model by challenging the model to differentiate between relevant and irrelevant examples. Common strategies include:

- **Random Negatives:** Efficient but lacking semantic contrast, leading to suboptimal performance (Karpukhin et al., 2020).
- **BM25 Negatives:** Leverage lexical similarity, but often introduce biases, particularly in semantically rich domains (Robertson and Walker, 1994).
- **In-Batch Negatives:** Computationally efficient but limited to local semantic contrasts, often underperforming in dense retrieval tasks (Xiong et al., 2020).

Our framework complements these approaches by dynamically generating negatives that balance semantic similarity and contextual irrelevance, avoiding the pitfalls of static or random methods.

### 2.3 Domain-Specific Retrieval Challenges

Enterprise retrieval systems face unique challenges, such as ambiguous terminology, overlapping concepts, and private datasets (Meghwani, 2024). General-purpose methods such as BM25 or dense retrieval models (Qu et al., 2020) fail to capture domain-specific complexities effectively. Our approach addresses these gaps by curating hard negatives that align with enterprise-specific semantics, improving retrieval precision and robustness for proprietary datasets.

We further discuss negative sampling techniques in Appendix A.1.

## 3 Methodology

To effectively train and finetune reranker models for domain-specific retrieval, it is essential to systematically handle technical ambiguities stemming from specialized terminologies, overlapping concepts, and abbreviations prevalent within enterprise domains.

We propose a structured, modular framework that integrates diverse embedding models, dimensionality reduction, and a novel semantic criterion for hard-negative selection. Figure 1 illustrates the high-level pipeline, components and their interactions. The re-ranking models fine-tuned using the

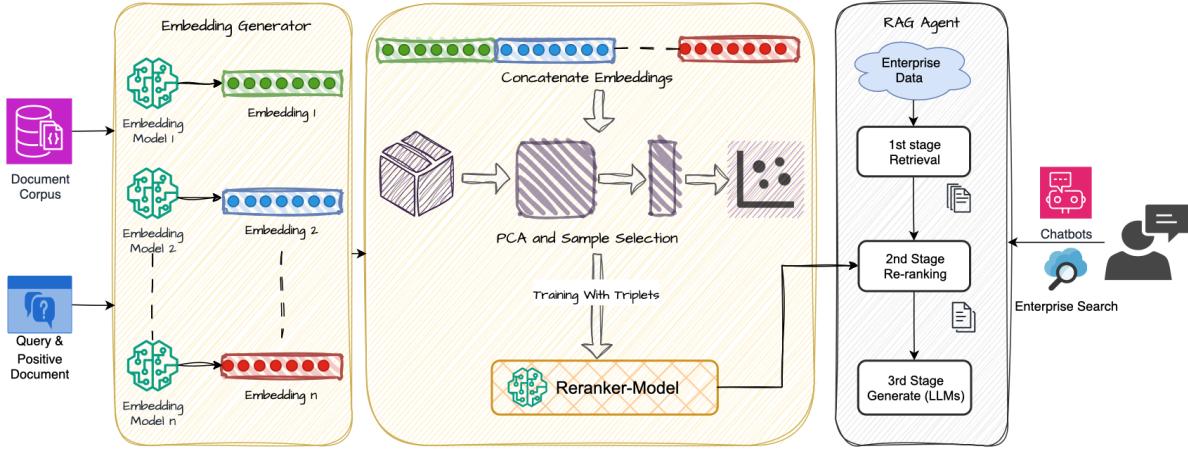


Figure 1: Overview of the methodology pipeline for training reranker models, including embedding generation, PCA-based dimensionality reduction and hard negative selection for fine-tuning.

hard negatives generated by our framework are directly deployed in downstream applications, such as RAG, significantly improving the resolution of customer queries through enhanced retrieval.

Our approach begins by encoding queries and documents into semantically rich vector representations using an ensemble of state-of-the-art bi-encoder embedding models. These embeddings are strategically selected based on multilingual support, embedding quality, training data diversity, context length handling, and performance (details provided in Appendix A.2). To manage embedding dimensionality and improve computational efficiency, Principal Component Analysis (PCA) (Maćkiewicz and Ratajczak, 1993) is utilized to project the concatenated embeddings onto a lower-dimensional space, maintaining 95% of the original variance.

We then define two semantic conditions (Eq. 5 and Eq. 6) to dynamically select high-quality hard negatives, addressing semantic similarity challenges and minimizing false negatives. Together, these two equations ensure that the selected hard negative is not only close to the query (Eq. 5) but also contextually distinct from the true positive, minimizing the risk of selecting topic duplicates or noisy positives (Eq. 6). For example, a query about deploying MySQL on Oracle Cloud, PD is a guide on that topic, and D is a doc about MySQL on-premise — semantically close to Q, but distant from PD.

Below we detail each methodological component, emphasizing their contributions to enhancing retrieval precision in domain-specific or enterprise retrieval tasks.

	Total	Train	Test
$< Q, PD >$	5250	1000	4250

Table 1: Dataset distribution of queries (Q) and positive documents (PD).

### 3.1 Dataset Statistics

Our experiments leverage a proprietary corpus containing 36,871 unannotated documents sourced from over 30 enterprise cloud services. Additionally, we prepared 5250 annotated query-positive document pairs ( $< Q, PD >$ ) for training and testing. Notably, we adopted a non-standard train-test split (as summarized in Table 1), allocating four times more data to testing than training to rigorously evaluate model robustness against varying training data volumes (additional analyses in Appendix A.4). To further validate generalizability, we conduct evaluations on publicly available domain-specific benchmarks: FiQA (finance) (TheFinAI, 2018), Climate Fever (climate science) (Diggelmann et al., 2021), and TechQA (technology) (Castelli et al., 2019). Detailed dataset statistics are provided in Appendix A.2.1.

### 3.2 Embedding Generation

Embeddings for queries, positive documents, and the corpus are computed via six diverse, high-performance bi-encoder models  $E_1, E_2, \dots, E_6$ , each selected strategically for capturing complementary semantic perspectives:

$$\mathbf{E}_k(x) \in \mathbb{R}^{d_k} \quad (1)$$

where  $d_k$  is the embedding dimension of the  $k_{th}$  model for textual input  $x$ . Concatenation of these

embeddings yields a comprehensive representation:

$$\mathbf{X}_{\text{concat}} = [\mathbf{e}_1(x); \mathbf{e}_2(x); \dots; \mathbf{e}_6(x)] \quad (2)$$

where  $\mathbf{X}_{\text{concat}} \in \mathbb{R}^{\sum_{k=1}^6 d_k}$  represents the concatenated embedding for the input  $x$ .

### 3.3 Dimensionality Reduction

To alleviate the computational overhead arising from high-dimensional concatenated embeddings, we apply PCA to reduce dimensionality while preserving semantic richness:

$$\mathbf{X}_{\text{PCA}} = \mathbf{X}_{\text{concat}} \mathbf{P}, \quad (3)$$

where  $\mathbf{P}$  represents the PCA projection matrix. We specifically select PCA due to its computational efficiency, and scalability, essential given our large enterprise corpus and high-dimensional embedding space. While we empirically evaluated nonlinear dimensionality reduction methods such as UMAP (McInnes et al., 2020) and t-SNE (Van der Maaten and Hinton, 2008), they offered negligible performance improvements over PCA but incurred substantially higher computational costs, making them impractical for deployment at scale in enterprise systems.

### 3.4 Hard Negative Selection Criteria

We propose two semantic criteria to identify high-quality hard negatives. PCA-reduced embeddings  $\mathbf{X}_{\text{PCA}}$  are organized around each query  $Q$ . For each query-positive document pair  $(Q, PD)$ , candidate documents  $D$  from the corpus are evaluated via cosine distances:

$$d(Q, PD), \quad d(Q, D), \quad d(PD, D) \quad (4)$$

A document  $D$  is selected as a hard negative only if it satisfies both criteria:

$$d(Q, D) < d(Q, PD) \quad (5)$$

$$d(Q, D) < d(PD, D) \quad (6)$$

Equation (5) ensures that the candidate negative document is semantically closer to the query than the actual positive document, making it a challenging negative example that potentially confuses the reranking model. Equation (6), ensures that the selected hard negative is not just query-confusing but also sufficiently dissimilar from the actual positive (avoiding near-duplicates or false negatives).

The candidate document  $D_{HN}$  with minimal  $d(Q, D)$  satisfying these conditions is chosen as

the primary hard negative. Additional hard negatives can similarly be selected based on semantic proximity rankings.

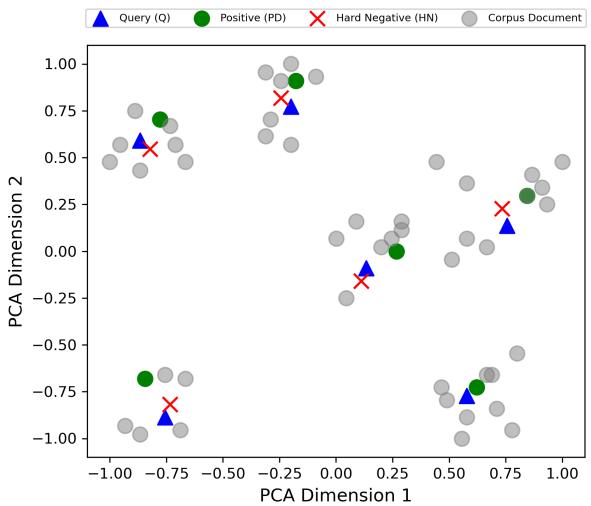


Figure 2: Hard negative selection on the first two PCA components (78% variance).  $Q$  act as centroids,  $PD$  guide selection of hard negatives; which are chosen based on semantic proximity.

Figure 2 illustrates an example embedding space, clearly depicting the query  $Q$ , positive document  $PD$ , and selected hard negative  $D_{HN}$ , visualizing the semantic selection criteria. In cases where no documents satisfy these conditions, no hard negatives are selected for that particular query. Further details on our embedding model & fine-tuning using these hard negatives are provided in Appendix A.2.

## 4 Experiments & Results

To evaluate the effectiveness of our proposed hard-negative selection framework, we conduct extensive experiments on our internal cloud-specific enterprise dataset, as well as domain-specific open-source benchmarks. We systematically compare our approach against multiple competitive negative sampling methods and perform detailed ablation studies to understand the contribution of individual framework components. Complete details on experimental setups and hyperparameters are provided in Appendix A.3.

### 4.1 Results & Discussion

**Comparative Analysis of Negative Sampling Strategies** Table 3 presents a detailed comparison of our negative sampling technique against several established methods, including Random, BM25, In-batch, STAR, and ADORE+STAR. The

Re-ranker (Fine-tuned w/)	Internal		FiQA		Climate-FEVER		TechQA	
	MRR@3	MRR@10	MRR@3	MRR@10	MRR@3	MRR@10	MRR@3	MRR@10
Baseline (No Fine-tuning)	0.42	0.45	0.45	0.48	0.44	0.46	0.57	0.61
In-batch Negatives	0.47	0.52	0.46	0.52	0.44	0.47	0.57	0.62
STAR	0.53	0.56	0.51	0.54	0.47	0.49	0.61	0.63
ADORE+STAR	0.54	0.57	0.52	0.54	0.48	0.52	0.63	0.66
<b>Our Proposed HN</b>	<b>0.57</b>	<b>0.64</b>	<b>0.54</b>	<b>0.56</b>	<b>0.52</b>	<b>0.55</b>	<b>0.65</b>	<b>0.69</b>

Table 2: Comparative performance benchmarking of our in-house reranker across multiple domain-specific datasets. The reranker is fine-tuned (FT) with different negative sampling techniques, highlighting the effectiveness of our proposed hard-negative mining method (HN).

Negative Sampling Method	MRR@3	MRR@10
Baseline	0.42	0.45
FT with Random Neg	0.47	0.51
FT with BM25 Neg	0.49	0.54
FT with In-batch Neg	0.47	0.52
FT with BM25+In-batch Neg	0.52	0.54
FT with STAR	0.53	0.56
FT with ADORE+STAR	0.54	0.57
FT with our HN	<b>0.57</b>	<b>0.64</b>

Table 3: Comparison of negative sampling methods for fine-tuning(FT) in-house cross-encoder reranker model. The proposed framework achieves 15% and 19% improvements in MRR@3 and MRR@10, respectively, over baseline methods.

baseline is defined as the performance of our internal reranker model without any fine-tuning. Our method achieves notable relative improvements of 15% in MRR@3 and 19% in MRR@10 over this baseline. The semantic nature of our hard negatives allows the reranker to distinguish contextually irrelevant but semantically similar documents effectively. In contrast, simpler baselines like Random or BM25 negatives suffer due to no semantic consideration, while advanced methods like STAR and ADORE+STAR occasionally miss subtle semantic nuances that our formulated selection criteria address effectively.

**Generalization Across Open-source Models** To validate the robustness and versatility of our framework, we evaluated various open-source embedding and reranker models (Table 4), clearly demonstrating improvements across all models when fine-tuned using our proposed negative sampling compared to ADORE+STAR and baseline (no fine-tuning). Notably, rerankers with multilingual capabilities, such as the BGE-Reranker and Jina Reranker, demonstrated pronounced improvements, likely benefiting from our embedding ensemble’s multilingual semantic richness. Similarly, larger models like e5-mistral exhibit significant gains, re-

flecting their capacity to exploit nuanced semantic differences provided by our negative samples. This analysis underscores the general applicability and model-agnostic benefits of our approach.

Model	Baseline	ADORE+STAR	Ours
Alibaba-NLP (gte-multilingual-reranker-base)	0.39	0.42	<b>0.45</b>
BGE-Reranker (bge-reranker-large)	0.44	0.47	<b>0.52</b>
Cohere Embed English Light (cohere-embed-english-light-v3.0)	0.32	0.34	<b>0.38</b>
Cohere Embed Multilingual (cohere-embed-multilingual-v3.0)	0.34	0.37	<b>0.40</b>
Cohere Reranker (rerank-multilingual-v2.0)	0.42	0.45	<b>0.49</b>
IBM Reranker (re2g-reranker-nq)	0.40	0.43	<b>0.46</b>
Inffloat Reranker (e5-mistral-7b-instruct)	0.35	0.38	<b>0.42</b>
Jina Reranker v2 (jina-reranker-v2-base-multilingual)	0.45	0.48	<b>0.53</b>
MS-MARCO (ms-marco-MiniLM-L-6-v2)	0.41	0.43	<b>0.46</b>
Nomic AI Embed Text (nomic-embed-text-v1.5)	0.33	0.36	<b>0.39</b>
NVIDIA			
NV-Embed-v2	0.38	0.41	<b>0.44</b>
Salesforce			
SFR-Embedding-2_R	0.37	0.40	<b>0.43</b>
Salesforce			
SFR-Embedding-Mistral	0.36	0.39	<b>0.42</b>
T5-Large	0.41	0.44	<b>0.47</b>

Table 4: Performance benchmarking (MRR@3) of reranker and embedding models using the proposed hard negative selection framework, compared with ADORE+STAR and baseline methods.

**Effectiveness on Domain-specific Public Datasets** We further tested our method’s adaptability across diverse public domain-specific datasets (FiQA, Climate-FEVER, TechQA), as shown in Table 2. Each dataset presents distinct retrieval challenges, ranging from technical jargon in TechQA to complex domain-specific reasoning in Climate-FEVER. Fine-tuning with our generated hard negatives consistently improved retrieval across these varied datasets. FiQA exhibited significant gains, likely due to the semantic differentiation required in finance-specific queries. These results demonstrate that our negative

sampling method is not only effective within our internal enterprise corpus but also valuable across diverse, domain-specific public datasets, indicating broad applicability and domain independence.

	<b>Model</b>	<b>MRR@3</b>	<b>MRR@10</b>
<b>Short Documents</b>	Baseline	0.481	0.526
	FT w/ proposed HN	<b>0.61</b>	<b>0.662</b>
<b>Long Documents</b>	Baseline	0.423	0.477
	FT w/ proposed HN	<b>0.475</b>	<b>0.521</b>

Table 5: Performance comparison of the in-house reranker without fine-tuning (Baseline) versus fine-tuned (FT) with our proposed hard negatives (HN), evaluated separately on short and long documents.

**Performance Analysis on Short vs. Long Documents** An explicit analysis of short versus long documents (Table 5) revealed differential performance gains. Short documents (under 1024 tokens) experienced substantial performance improvements (MRR@3 improving from 0.481 to 0.61), attributed to minimal semantic redundancy and tokenization constraints. Conversely, long documents showed more moderate improvements (MRR@3 from 0.423 to 0.475), primarily due to embedding truncation that causes loss of context and increased semantic complexity. Future research should focus explicitly on developing hierarchical or segment-based embedding methods to address these limitations.

**Ablation Studies** To clearly understand the impact of the individual components of the framework, we conducted systematic ablation studies (Table 6). Training with positive documents alone produced only slight gains (+0.03 MRR@3), reaffirming the critical role of high-quality hard negatives. Evaluating individual embedding models separately indicated varying performance due to their differing semantic representations and underlying training. However, the concatenation of diverse embeddings provided significant performance improvements (+0.15 MRR@3), clearly highlighting the advantages of capturing semantic diversity.

Additionally, PCA-based dimensionality reduction analysis identified the optimal variance threshold at 95%. Lower thresholds resulted in marked semantic degradation, reducing retrieval performance. This trade-off highlights PCA as an essential efficiency-enhancing step for the framework.

Collectively, these detailed analyses underscore our method’s strengths, limitations, and method-

ological rationale, providing clear empirical justification for each design decision.

#	<b>Proposed Strategies</b>	<b>MRR@3</b>	<b>MRR@10</b>
1	Baseline	0.42	0.45
<b>Positive Document (PD) Only</b>			
2	Fine-tuning with PD Only	0.45	0.51
<b>Hard Negative(HN) with Embedding <math>E_k</math></b>			
3a	HN with $E_1 + \text{PD}$	0.45	0.51
3b	HN with $E_2 + \text{PD}$	0.47	0.53
3c	HN with $E_3 + \text{PD}$	0.51	0.55
3d	HN with $E_4 + \text{PD}$	0.45	0.52
3e	HN with $E_5 + \text{PD}$	0.48	0.51
3f	HN with $E_6 + \text{PD}$	0.49	0.52
3g	HN with $X_{\text{concat}} + \text{PD}$	<b>0.57</b>	<b>0.64</b>
<b>XPCA Variance Impact + PD</b>			
4a	HN with XPCA (99% Variance)	0.57	0.64
4b	HN with XPCA (95% Variance)	<b>0.57</b>	<b>0.64</b>
4c	HN with XPCA (90% Variance)	0.55	0.63
4d	HN with XPCA (80% Variance)	0.51	0.58
4e	HN with XPCA (70% Variance)	0.49	0.56

Table 6: Results of ablation study showing the impact of embeddings, PCA variance thresholds, and positive documents on MRR, on the in-house re-ranker model.

## 4.2 Case Studies: Examples of Hard Negative Impact

Figure 3 shows how similar topics in the domain of cloud computing. To demonstrate the qualitative benefits of the proposed framework, we present two case studies where the baseline and fine-tuned models produce different ranking results. These examples highlight the significance of hard negatives in distinguishing semantically similar but contextually irrelevant documents.

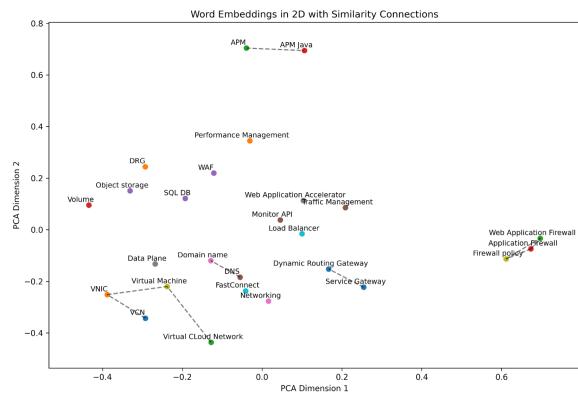


Figure 3: Illustrations of similar topics in the domain of Cloud Computing

### Case Study 1: Disambiguating Technical Acronyms.

- **Query (Q):** "What is VCN in Cloud Infrastructure?"

- **Positive Document (PD):** A document explaining "Virtual Cloud Network (VCN)" in Cloud Infrastructure, detailing its setup and usage.
- **Hard Negative (HN):** A document discussing "Virtual Network Interface Card (VNIC)" in the context of networking hardware.

**Baseline Result:** The baseline model incorrectly ranks the hard negative above the positive document due to overlapping terms such as "virtual" and "network."

**Proposed Method Result:** The fine-tuned model ranks the positive document higher, correctly identifying the contextual match between the query and the description of VCN. This improvement is attributed to the triplet loss training with hard negatives.

### Case Study 2: Domain-Specific Terminology.

- **Query (Q):** "How does the CI WAF handle incoming traffic?"
- **Positive Document (PD):** A document explaining the Web Application Firewall (WAF) in CI, its configuration, and traffic filtering mechanisms.
- **Hard Negative (HN):** A document discussing general firewall configurations in networking.

**Baseline Result:** The baseline model ranks the hard negative higher due to lexical overlap between the terms "firewall" and "traffic."

**Proposed Method Result:** The proposed framework ranks the positive document higher, leveraging domain-specific semantic representations.

These case studies illustrate the practical advantages of training with hard negatives, especially in domains with overlapping terminology or acronyms.

Additional detailed analyses, illustrative practical implications for enterprise applications, and explicit future directions are discussed in detail in [A.4](#), and [A.5](#).

## 5 Conclusion

We introduced a scalable, modular framework leveraging dynamic ensemble-based hard-negative mining to significantly enhance re-ranking models in enterprise and domain-specific retrieval scenarios.

Our method dynamically curates semantically challenging yet contextually irrelevant negatives, allowing re-ranking models to effectively discriminate subtle semantic differences. Empirical evaluations on proprietary enterprise data and diverse public domain-specific benchmarks demonstrated substantial improvements of up to 15% in MRR@3 and 19% in MRR@10 over state-of-the-art negative sampling techniques, including BM25, In-Batch Negatives, STAR, and ADORE+STAR.

Our approach offers clear practical benefits in real-world deployments, benefiting downstream applications such as knowledge management, customer support systems, and Retrieval-Augmented Generation (RAG), where retrieval precision directly influences user satisfaction and Generative AI effectiveness. The strong performance and generalizability across various domains further underscore the framework's readiness for industry-scale deployment.

Future work will focus on extending our framework to handle incremental updates of enterprise knowledge bases and exploring real-time negative sampling strategies for continuously evolving corpora, further enhancing the adaptability and robustness required in practical industry settings.

## 6 Limitations

While our approach advances the state of hard negative mining and encoder-based retrieval, several limitations remain that open avenues for future research. One key challenge is the performance disparity between short and long documents. Addressing this requires more effective document chunking strategies and the development of hierarchical representations to preserve context across segments. Additionally, the retrieval of long documents is complicated by semantic redundancy and truncation, warranting deeper analysis of their structural complexity. Our current use of embedding concatenation for ensembling could also be refined—future work should evaluate alternative fusion techniques such as weighted averaging or attention-based mechanisms. Moreover, extending the retrieval framework to support cross-lingual and multilingual scenarios would enhance its utility in globally distributed applications.

## References

- AMIT AGARWAL. 2021. Evaluate generalisation & robustness of visual features from images to video.

- ResearchGate. Available at <https://doi.org/10.13140/RG.2.2.33887.53928>.
- Amit Agarwal, Srikant Panda, and Kulbhushan Pachauri. 2024a. Synthetic document generation pipeline for training artificial intelligence models. US Patent App. 17/994,712.
- Amit Agarwal, Srikant Panda, and Kulbhushan Pachauri. 2025. **FS-DAG**: Few shot domain adapting graph networks for visually rich document understanding. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 100–114, Abu Dhabi, UAE. Association for Computational Linguistics.
- Amit Agarwal, Hitesh Patel, Priyaranjan Pattnayak, Srikant Panda, Bhargava Kumar, and Tejaswini Kumar. 2024b. Enhancing document ai data generation through graph-based synthetic layouts. *arXiv preprint arXiv:2412.03590*.
- Jina AI. 2023. **jina-reranker-v2-base-multilingual**.
- Arian Askari, Mohammad Aliannejadi, Evangelos Kanoulas, and Suzan Verberne. 2023. Generating synthetic documents for cross-encoder re-rankers: A comparative study of chatgpt and human experts.
- Jiaqi Bai, Hongcheng Guo, Jiaheng Liu, Jian Yang, Xinnian Liang, Zhao Yan, and Zhoujun Li. 2023. **Griprank: Bridging the gap between retrieval and generation via the generative knowledge improved passage ranking**. *Preprint*, arXiv:2305.18144.
- Vittorio Castelli, Rishav Chakravarti, Saswati Dana, Anthony Ferritto, Radu Florian, Martin Franz, Dinesh Garg, Dinesh Khandelwal, Scott McCarley, Mike McCawley, Mohamed Nasr, Lin Pan, Cezar Pendus, John Pitrelli, Saurabh Pujar, Salim Roukos, Andrzej Sakrajda, Avirup Sil, Rosario Uceda-Sosa, Todd Ward, and Rong Zhang. 2019. The techqa dataset. *Preprint*, arXiv:1911.02984.
- Cohere. 2023a. Cohere-embed-multilingual-v3.0. Available at: <https://cohere.com/blog/introducing-embed-v3>.
- Cohere. 2023b. Reranker model. Available at: <https://docs.cohere.com/v2/docs/reranking-with-cohere>.
- Gabriel de Souza P. Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. 2024. **Nv-retriever: Improving text embedding models with effective hard-negative mining**. *Preprint*, arXiv:2407.15831.
- Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bülian, Massimiliano Ciaramita, and Markus Leipold. 2021. **Climate-fever: A dataset for verification of real-world climate claims**. *Preprint*, arXiv:2012.00614.
- Karan Dua, Praneet Pabolu, and Mengqing Guo. 2024. Generating templates for use in synthetic document generation processes. US Patent App. 18/295,765.
- Karan Dua, Praneet Pabolu, and Ranjeet Kumar Gupta. 2025. Generation of synthetic doctor-patient conversations. US Patent App. 18/495,966.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. **Language-agnostic bert sentence embedding**. *Preprint*, arXiv:2007.01852.
- Luyu Gao and Jamie Callan. 2021. **Condenser: a pre-training architecture for dense retrieval**. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 981–993.
- Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfonso Gliozzo. 2022. **Re2G: Retrieve, rerank, generate**. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715, Seattle, United States. Association for Computational Linguistics.
- Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. **How close is chatgpt to human experts? comparison corpus, evaluation, and detection**.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. **Realm: Retrieval-augmented language model pre-training**.
- EK Jasila, N Saleena, and KA Abdul Nazeer. 2023. An efficient document clustering approach for devising semantic clusters. *Cybernetics and Systems*, pages 1–18.
- Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen Tau Yih. 2020. **Dense passage retrieval for open-domain question answering**. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 6769–6781.
- Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. **Nv-embed: Improved techniques for training llms as generalist embedding models**. *arXiv preprint arXiv:2405.17428*.
- Fulu Li, Zhiwen Xie, and Guangyou Zhou. 2024. **Theme-enhanced hard negative sample mining for open-domain question answering**. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12436–12440.
- Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*.
- Ye Liu, Kazuma Hashimoto, Yingbo Zhou, Semih Yavuz, Caiming Xiong, and Philip S. Yu. 2021. **Dense hierarchical retrieval for open-domain question answering**. In *Conference on Empirical Methods in Natural Language Processing*.

- Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. 2019. **Cedr: Contextualized embeddings for document ranking.** *SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1101–1104.
- Andrzej Maćkiewicz and Waldemar Ratajczak. 1993. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342.
- Leland McInnes, John Healy, and James Melville. 2020. **Umap: Uniform manifold approximation and projection for dimension reduction.** *Preprint*, arXiv:1802.03426.
- Hansa Meghwani. 2024. **Enhancing retrieval performance: An ensemble approach for hard negative mining.** *Preprint*, arXiv:2411.02404.
- Vivek Mehta, Mohit Agarwal, and Rohit Kumar Kaliyar. 2024. A comprehensive and analytical review of text clustering techniques. *International Journal of Data Science and Analytics*, pages 1–20.
- Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024a. **Sfr-embedding-2: Advanced text embedding with multi-stage training.**
- Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024b. **Sfr-embedding-mistral: Enhance text retrieval with transfer learning.** Salesforce AI Research Blog.
- Thanh-Do Nguyen, Chi Minh Bui, Thi-Hai-Yen Vuong, and Xuan-Hieu Phan. 2022. Passage-based bm25 hard negatives: A simple and effective negative sampling strategy for dense retrieval.
- Rodrigo Nogueira and Kyunghyun Cho. 2019. **Passage re-ranking with bert.**
- Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-stage document ranking with bert.
- Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. **Nomic embed: Training a reproducible long context text embedder.** *Preprint*, arXiv:2402.01613.
- Praneet Paboli, Karan Dua, and Sriram Chaudhury. 2024a. Multi-lingual natural language generation. US Patent App. 18/318,315.
- Praneet Paboli, Karan Dua, and Sriram Chaudhury. 2024b. Multi-lingual natural language generation. US Patent App. 18/318,327.
- Srikant Panda, Amit Agarwal, Goutham Nambirajan, and Kulbhushan Pachauri. 2025a. Out of distribution element detection for information extraction. US Patent App. 18/347,983.
- Srikant Panda, Amit Agarwal, and Kulbhushan Pachauri. 2025b. Techniques of information extraction for selection marks. US Patent App. 18/240,344.
- Hitesh Laxmichand Patel, Amit Agarwal, Arion Das, Bhargava Kumar, Srikant Panda, Priyanjan Pattnayak, Taki Hasan Rafi, Tejaswini Kumar, and Dong-Kyu Chae. 2025. Sweeval: Do llms really swear? a safety benchmark for testing limits for enterprise use. 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 558–582.
- Hitesh Laxmichand Patel, Amit Agarwal, Bhargava Kumar, Karan Gupta, and Priyanjan Pattnayak. 2024. Llm for barcodes: Generating diverse synthetic data for identity documents. *arXiv preprint arXiv:2411.14962*.
- Priyanjan Pattnayak, Amit Agarwal, Hansa Meghwani, Hitesh Laxmichand Patel, and Srikant Panda. 2025a. Hybrid ai for responsive multi-turn online conversations with novel dynamic routing and feedback adaptation. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 215–229.
- Priyanjan Pattnayak, Hitesh Laxmichand Patel, and Amit Agarwal. 2025b. **Tokenization matters: Improving zero-shot ner for indic languages.** *Preprint*, arXiv:2504.16977.
- Priyanjan Pattnayak, Hitesh Laxmichand Patel, Amit Agarwal, Bhargava Kumar, Srikant Panda, and Tejaswini Kumar. 2025c. **Clinical qa 2.0: Multi-task learning for answer extraction and categorization.** *Preprint*, arXiv:2502.13108.
- Ronak Pradeep, Yuqi Liu, Xinyu Zhang, Yilin Li, Andrew Yates, and Jimmy Lin. 2022. **Squeezing water from a stone: A bag of tricks for further improving cross-encoder effectiveness for reranking.** In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 13185 LNCS, pages 655–670. Springer Science and Business Media Deutschland GmbH.
- Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2020. **Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering.**
- 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.
- Nils Reimers and Iryna Gurevych. 2019. **Sentence-bert: Sentence embeddings using siamese bert-networks.**

- In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.
- S. E. Robertson and S. Walker. 1994. [Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval](#), pages 232–241. Springer London.
- Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. [jina-embeddings-v3: Multilingual embeddings with task lora](#). Preprint, arXiv:2409.10173.
- TheFinAI. 2018. [Fifa: A financial question answering dataset](#). Available at Hugging Face.
- Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. [Journal of machine learning research](#), 9(11).
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. arXiv preprint arXiv:2401.00368.
- Svante Wold, Kim H. Esbensen, Kim H. Esbensen, Paul Geladi, and Paul Geladi. 1987. [Principal component analysis](#). [Chemometrics and Intelligent Laboratory Systems](#), 2:37–52.
- Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. [C-pack: Packaged resources to advance general chinese embedding](#). Preprint, arXiv:2309.07597.
- Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. [Approximate nearest neighbor negative contrastive learning for dense text retrieval](#).
- Zhen Yang, Zhou Shao, Yuxiao Dong, and Jie Tang. 2024. [Trisampler: A better negative sampling principle for dense retrieval](#). Preprint, arXiv:2402.11855.
- Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. [Optimizing dense retrieval model training with hard negatives](#). [SIGIR 2021 - Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval](#), pages 1503–1512.
- Dun Zhang. 2024. [stella-embedding-model-2024](#).
- Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. arXiv preprint arXiv:2407.19669.

## A Appendix

### A.1 Extended Related Work

**Hard Negatives in Retrieval Models** Static and dynamic hard negatives have been extensively studied. Static negatives, such as those generated by BM25 (Robertson and Walker, 1994) or PassageBM25 (Nguyen et al., 2022), provide challenging lexical contrasts but risk overfitting due to their fixed nature (Qu et al., 2020). Dynamic negatives, as introduced in ANCE (Xiong et al., 2020) and ADORE (Zhan et al., 2021) adapt during training, other effective methods like positive-aware mining (de Souza P. Moreira et al., 2024), theme-enhanced negatives (Li et al., 2024) offers relevant challenges but incurring high computational costs due to periodic re-indexing and bigger embedding dimension. Our framework mitigates these issues by leveraging clustering and dimensionality reduction to dynamically identify negatives without requiring re-indexing.

Localized Contrastive Estimation (LCE) (Guo et al., 2023; AGARWAL, 2021) further demonstrated the effectiveness of incorporating hard negatives into cross-encoder training, improving reranking accuracy when negatives align with retriever outputs. Additionally, (Pradeep et al., 2022) highlighted the importance of hard negatives even in advanced pretraining setups like Condenser (Gao and Callan, 2021), which emphasizes their necessity for robust optimization.

**Advances in Dense Retrieval and Cross-Encoders** Dense retrieval models like DPR (Karpukhin et al., 2020) and REALM (Guu et al., 2020) encode queries and documents into dense embeddings, enabling semantic matching. Recent advances in dense retrieval and ranking include GripRank’s generative knowledge-driven passage ranking (Bai et al., 2023), Dense Hierarchical Retrieval’s multi-stage framework for efficient question answering (Liu et al., 2021; Pattnayak et al., 2025a,c,b; Patel et al., 2025), and TriSampler’s optimized negative sampling for dense retrieval (Yang et al., 2024), collectively enhancing retrieval performance. Cross-encoders, such as monoBERT (Nogueira et al., 2019; Nogueira and Cho, 2019), further improve retrieval precision by jointly encoding query-document pairs but require high-quality training data, particularly challenging negatives (MacAvaney et al., 2019; Panda et al., 2025b). Techniques such

as synthetic data generation (Askari et al., 2023; Agarwal et al., 2024a, 2025) augment training datasets but lack the realism and semantic depth provided by our hard negative mining approach.

**Dimensionality Reduction in IR** Clustering methods have been used to group semantically similar documents, improving retrieval efficiency and training data organization (Mehta et al., 2024; Jasila et al., 2023; Dua et al., 2025; Panda et al., 2025a). Dimensionality reduction techniques like PCA (Wold et al., 1987) enhance scalability by reducing computational complexity. Our framework uniquely combines these techniques to dynamically identify negatives that challenge retrieval models in a scalable manner.

**Synthetic Data in Retrieval** Recent work (Askari et al., 2023; Agarwal et al., 2024a,b; Patel et al., 2024; Dua et al., 2024; Pabolu et al., 2024a,b) has explored using large language models to generate synthetic training data for retrieval tasks. While effective in low-resource settings, synthetic data often struggles with factual inaccuracies and domain-specific relevance. In contrast, our framework relies on real-world data to curate semantically challenging negatives, ensuring high-quality training samples without introducing synthetic biases.

**Summary of Contributions** While previous works address various aspects of negative sampling, hard negatives, and synthetic data, our approach bridges the gap between static and dynamic strategies. By dynamically curating negatives using clustering and dimensionality reduction, we achieve a scalable and semantically precise methodology tailored to domain-specific retrieval tasks.

## A.2 Extended Methodology

### A.2.1 Dataset Statistics

**Queries Length Distribution** In this section we analyze the distribution of queries length in our enterprise dataset. Figure 4 shows that the length of queries ranges from 1 to 25 words, with some queries having very few words. This highlights that user queries can sometime be just 2-3 words about a topic, increasing the probability of retrieving documents mentioning those topics or concepts which can be contextually different. Therefore, when we select hard negatives, it is crucial to consider not only the relationship between the query and documents but also the relationship between the

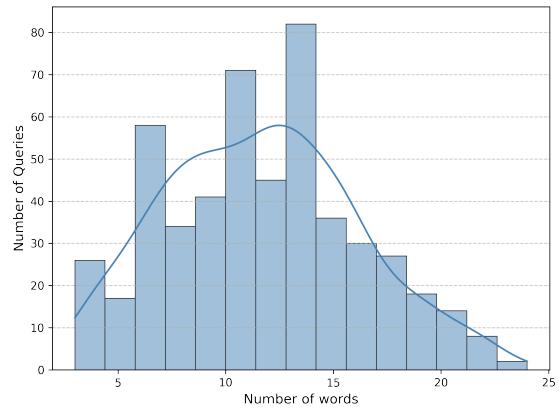


Figure 4: Length Distribution of queries in the dataset.

positive document and other documents, ensuring a comparison with texts on similar topics and similar lengths.

Model ( $E_k$ )	Params (M)	Dimension	Max Tokens
stella_en_400M_v5	435	8192	8192
jina-embeddings-v3 (multilingual)	572	1024	8194
mxbai-embed-large-v1	335	1024	512
bge-large-en-v1.5	335	1024	512
LaBSE (multilingual)	471	768	256
all-mpnet-base-v2 (multilingual)	110	768	514

Table 7: Embedding models used to construct  $X_{\text{concat}}$ , combining diverse semantic representations for queries ( $Q$ ), positive documents ( $PD$ ), and corpus documents ( $D$ ).

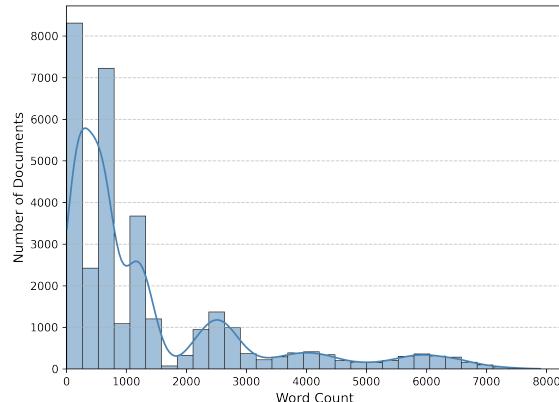


Figure 5: Shows document length distribution in Enterprise corpus.

**Document Length Distribution** As shown in Figure 5, document lengths are significantly longer than query lengths. This disparity in context length affects the similarity scores, potentially reducing the accuracy of retrieval systems. In our in-house dataset, each query is paired with a single correct document (though its not limited by number of positive-negative document per query). This positive document is crucial for identifying challenging hard negatives and hence helpful for encoder-based model training.

### A.2.2 Embedding Models

Table 7 lists the embedding models (Zhang, 2024; Sturua et al., 2024; Li and Li, 2023; Xiao et al., 2023; Feng et al., 2022; Reimers and Gurevych, 2019; Zhang et al., 2024) used to construct  $X_{\text{concat}}$ , combining diverse semantic representations for queries ( $Q$ ), positive documents ( $PD$ ), and corpus documents ( $D$ ). These models were selected for their performance, model size, ability to handle multilingual context, providing complementary strengths in dimensionality and token coverage. By integrating embeddings from these models, the framework captures nuanced semantic relationships crucial for reranker training.

### A.2.3 Unified Contrastive Loss

The unified contrastive loss is designed to improve ranking precision for both bi-encoders and cross-encoders, by ensuring that positive documents ( $PD$ ) are ranked closer to the query ( $Q$ ) than hard negatives ( $D_{HN}$ ) by a margin  $m$ . The loss is defined as:

$$L = \sum_{i=1}^N \max(0, m + d(Q_i, PD_i) - d(Q_i, D_{HN_i})) \quad (7)$$

where:

- $PD_i$ : Positive document associated with query  $Q_i$ .
- $D_{HN_i}$ : Hard negative document, semantically similar to  $PD_i$  but contextually irrelevant.
- $d(Q_i, D_i)$ : Distance metric measuring relevance between  $Q_i$  and  $D_i$ .
- $m$ : Margin ensuring  $PD_i$  is closer to  $Q_i$  than  $D_{HN_i}$  by at least  $m$ , encouraging the model to distinguish between relevant and irrelevant documents effectively.

For **bi-encoders**, the distance metric is defined as:

$$d(Q_i, D_i) = 1 - \cosine(e_{Q_i}, e_{D_i}), \quad (8)$$

where  $e_{Q_i}$  and  $e_{D_i}$  are the embeddings of the query and document, respectively. For **cross-encoders**, the distance metric is:

$$d(Q_i, D_i) = -s(Q_i, D_i), \quad (9)$$

where  $s(Q_i, D_i)$  is the cross-encoder’s relevance score for the query-document pair.

This formulation leverages the triplet of  $(Q, PD, D_{HN})$  to minimize  $d(Q_i, PD_i)$ , pulling positive documents closer to the query, while maximizing  $d(Q_i, D_{HN_i})$ , pushing hard negatives further away. By emphasizing semantically challenging examples, the model learns sharper decision boundaries for improved ranking precision.

### A.3 Experimental Setup

**Datasets** We evaluate our framework extensively using both proprietary and public datasets:

- **Internal Proprietary Dataset:** Consisting of approximately 5250 query-document pairs, on cloud services like computing, networking, firewall, ai services. It includes both short ( $< [1024 \text{ tokens}]$ ) and long documents ( $\geq [1024 \text{ tokens}]$ ).
- **FiQA Dataset:** A financial domain-specific dataset widely used for retrieval benchmarking.
- **Climate-FEVER Dataset:** An environment-specific fact-checking dataset focused on climate-related information retrieval.
- **TechQA Dataset:** A technical question-answering dataset emphasizing software engineering and technology-related queries.

**Training and Fine-tuning** All re-ranking models are fine-tuned using a triplet loss with margin with same hyper-parameters. Early stopping is employed based on validation MRR@10 scores to prevent overfitting.

**Evaluation Metrics** Model performance is evaluated using standard retrieval metrics: Mean Reciprocal Rank (MRR) at positions 3 and 10 (MRR@3 and MRR@10), which measure retrieval quality and ranking precision. Each reported metric is averaged across three experimental runs for robustness.

Strategy	Training Data	MRR@3	MRR@10
Baseline	0	0.42	0.45
	100	0.46	0.49
	200	0.48	0.51
(Ours)	300	0.50	0.53
	400	0.52	0.56
	500	0.52	0.58
	600	0.54	0.60
	700	0.54	0.62
	800	0.56	0.63
	900	0.57	0.64
	1000	<b>0.57</b>	<b>0.64</b>

Table 8: Comparison of Strategies with Varying Training Data Sizes

#### A.4 Extended Results & Ablation

**Impact of Training Data Size** As shown in Table 8, both MRR@3 and MRR@10 improve as the training data size increases, with more pronounced gains in MRR@10. MRR@3 shows gradual improvement, from 0.42 at the baseline to 0.57 with 100 examples, highlighting the model’s enhanced ability to rank relevant documents within the top 3. MRR@10, on the other hand, shows more significant improvement, from 0.45 to 0.64, indicating that the model benefits more from additional data when considering the top 10 ranked documents.

Our method shows promising results even with smaller training sets, demonstrating the effectiveness of incorporating hard negatives early in the training process. This suggests that hard negatives significantly enhance the model’s ability to distinguish relevant from irrelevant documents against a given query, even when data is limited. This approach is particularly beneficial in enterprise contexts, where annotated data may be scarce, enabling quicker improvements in domain-specific retrieval performance.

**Models in the Study** In our study we compared the performance of other finetuned re-ranker (Glass et al., 2022; Wang et al., 2023; Raffel et al., 2020) and embedding models (Zhang et al., 2024; Nussbaum et al., 2024) using hard negatives generated by our proposed framework in Table 4. We benchmarked the BGE-Reranker (Xiao et al., 2023), NV-Embed (Lee et al., 2024) Salesforce-SFR (Meng et al., 2024a,b), jina-reranker (AI, 2023) and Cohere-Reranker (Cohere, 2023a,b),

##### A.4.1 Analysis of Long vs. Short Documents

Table 5 reveals a consistent disparity in MRR scores between short and long documents, with

long documents showing lower performance. Here, we analyze potential reasons and propose mitigation strategies.

##### Challenges with Long Documents.

- **Semantic Redundancy:** Long documents often contain repetitive or tangential content, diluting their relevance to a specific query.
- **Context Truncation:** Fixed-length tokenization (e.g., 512 or 1024 tokens) truncates long documents, potentially discarding critical information.
- **Query-to-Document Mismatch:** Short queries may not provide sufficient context to match the nuanced information spread across a lengthy document.

##### Potential Solutions.

- **Chunk-Based Retrieval:** Split long documents into smaller, semantically coherent chunks and rank them individually.
- **Hierarchical Embeddings:** Use hierarchical models to aggregate sentence- or paragraph-level embeddings for better context representation.
- **Query Expansion:** Enhance short queries with additional context using techniques like query rewriting or pseudo-relevance feedback.

This analysis highlights the need for future work to address the inherent challenges of ranking long documents effectively.

#### A.5 Practical Implications for Enterprise Applications

The proposed framework has significant practical implications for enterprise information retrieval systems, particularly in retrieval-augmented generation (RAG) pipelines.

**Improved Ranking Precision.** By training with hard negatives, the model ensures that the most relevant documents are retrieved for each query. This is particularly critical for enterprise use cases such as:

- **Technical Support:** Retrieving precise documentation for customer queries, reducing resolution times.
- **Knowledge Management:** Ensuring that employees access the most relevant internal resources quickly.

**Enhanced Generative Quality.** High-quality retrieval directly improves the factual accuracy and coherence of outputs generated by large language models in RAG pipelines. For example:

- **Documentation Summarization:** Summaries generated by models like GPT are more reliable when based on top-ranked, accurate sources.
- **Customer Interaction:** Chatbots generate more contextually relevant responses when fed precise retrieved documents.

**Scalability and Adaptability.** The framework's modular design, including the use of diverse embeddings and clustering-based hard negative selection, allows it to adapt to:

- Different industries (e.g., healthcare, finance, manufacturing).
- Multi-lingual or cross-lingual retrieval tasks.

These practical implications underscore the versatility and enterprise readiness of the proposed framework.