Hard Negatives, Hard Lessons: Revisiting Training Data Quality for Robust Information Retrieval with LLMs

Nandan Thakur* **Crystina Zhang*** **Xueguang Ma** **Jimmy Lin**

David R. Cheriton School of Computer Science,
University of Waterloo, Canada

Code: https://github.com/castorini/rlhn
Dataset: https://huggingface.co/rlhn

Abstract

Training robust retrieval and reranker models typically relies on large-scale retrieval datasets; for example, the BGE collection contains 1.6 million query-passage pairs sourced from various data sources. However, we find that certain datasets can negatively impact model effectiveness — pruning 8 out of 15 datasets from the BGE collection, reduces the training set size by 2.35×, surprisingly increases nDCG@10 on BEIR by 1.0 point. This motivates a deeper examination of training data quality, with a particular focus on "false negatives", where relevant passages are incorrectly labeled as irrelevant. We utilize LLMs as a simple, cost-effective approach to *identify* and *relabel* false negatives in training datasets. Experimental results show that relabeling false negatives as true positives improves both E5 (base) and Qwen2.5-7B retrieval models by 0.7–1.4 points on BEIR and by 1.7–1.8 points at nDCG@10 on zero-shot AIR-BENCH evaluation. Similar gains are observed for rerankers fine-tuned on the relabeled data, such as Qwen2.5-3B on BEIR. The reliability of LLMs to identify false negatives is supported by human annotation results. Our training dataset and code are publicly available.

1 Introduction

Modern-day retrievers and rerankers are data-hungry, relying on large and high-quality training datasets to accurately retrieve or rerank on challenging domains (Thakur et al., 2021; Muennighoff et al., 2023; Chen et al., 2025; Su et al., 2025). A typical training dataset for information retrieval (IR) has multiple instances consisting a training query, labeled positive passages, and a set of mined *hard negative passages*. Sampling hard negatives has been consistently used in retrieval models to improve downstream retrieval accuracy (Karpukhin et al., 2020; Xiong et al., 2021; Qu et al., 2021; Moreira et al., 2024, *inter alia*).

More recently, state-of-the-art (SoTA) retrieval models are observed to fine-tune on enormous or large training datasets (Zhang et al., 2025). While the general notion is that more training data is better, in accordance with scaling laws (Chen et al., 2024a; Li et al., 2024; Muennighoff et al., 2025), we show the contrary: fine-tuning on a select few datasets is rather crucial. For example, removing ELI5 surprisingly improves nDCG@10 on 7 out of 14 of the BEIR datasets (Thakur et al., 2021) and the average nDCG@10 by 0.6 points. A similar observation is also made on other training datasets: by pruning 8 out of the 15 datasets in the BGE training collection (Li et al., 2024),[1] the E5 (base) retrieval model improves by 1.0 point nDCG@10 on BEIR (as shown later in Figure 4).

The above observation reveals a non-negligible amount of "false" or mislabeled data is mixed in the current training datasets, that not only adds unnecessary training cost but also hurts the model training. *How can the "false" data be eliminated?* We approach the issue from the perspective of *false negatives* (example in Figure 1), specifically, by proposing **RLHN** (ReLabeling Hard Negatives) utilizing a cost-effective framework with large language model (LLM) cascading (Chen et al., 2024c) to accurately identify and relabel false negatives (at a data sample level). We choose to look at false negatives since it is a systematic pitfall from how retrieval training data is constructed:[2] As long as there are *unjudged* documents used as negative examples, the issue of false negative persists, which is especially severe for big sparsely-annotated datasets, such as MS MARCO (Nguyen et al., 2016) or NQ (Kwiatkowski et al., 2019).

The issue of false negatives has been noticed for long — Qu et al. (2021) distill knowledge from a

*Equal contribution.

[1] The pruned dataset contains only 42.5% training pairs of the original dataset, making it 2.35× smaller in size.

[2] In contrast to "false positives", that only results from mistakes of human annotators in training datasets.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 9064–9083
November 4-9, 2025

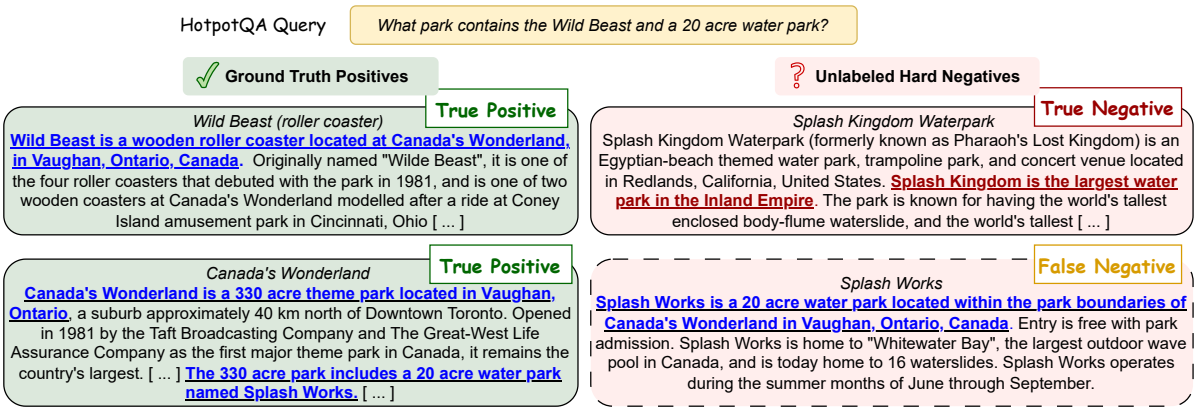

Figure 1: Example of a training instance (query, ground truth positives, and unlabeled hard negatives) with detected false negatives taken from HOTPOTQA. The false negative passage (*Splash Works*) is **mislabeled** as it is relevant in answering the user's query. The relevant parts of the text useful in answering the query are highlighted in **blue**.

cross-encoder to alleviate their impact. Moreira et al. (2024) filter potential false negatives based on relevance score to the query. However, the former solution does not curate or clean the training datasets and is based on the assumption that the cross-encoder is more robust to false negatives than retrieval models. As we will show in Section 5, albeit smaller, inferior training data also negatively affect cross-encoders. The latter solution is based on the assumption that the relevance scores of false negatives are systematically higher than 95% of the positive scores, which does not consider score variance at the level of a data instance.

We use an LLM cascading framework to alleviate "false negatives". The first stage employs GPT-4o-mini, a cost-effective LLM, to identify false negatives in all training instances. Next, the detected instances with false negatives are relabeled with a more reliable judge, GPT-4o. We observe a maximum of 56% of training pairs in MS MARCO can contain false negative documents, to a minimum of about 3% in SCIDOCSRR. The framework is better illustrated in Figure 2. With the false negatives detected, we compared three data modification approaches: (i) *remove*: discarding the whole training instance, (ii) *remove HN*: removing only the false hard negatives, and (iii) *relabel HN (RLHN)*: relabeling the false hard negatives as ground truth. We experiment on the seven pruned training datasets from the BGE training collection (Li et al., 2024).

Our results consistently show that the RLHN setting achieves the highest nDCG@10 scores on BEIR (Thakur et al., 2021) and AIR-BENCH (Chen et al., 2025), amongst their counterparts with both retrievers: E5 (base) and Qwen2.5-7B and a reranker with Qwen2.5-3B. Compared to the aforementioned works, RLHN outperforms hard negative sampling in Moreira et al. (2024) and is comparable to cross-encoder distillation (Qu et al., 2021) yet with a simpler training pipeline.

To better understand the behavior of LLM judgment in identifying false negatives, we compare LLM judgment with human assessors on 670 randomly sampled query–hard negative pairs. We observe the Cohen's Kappa (κ) score of GPT-4o is 10 points higher than GPT-4o-mini, which echoes their effectiveness in improving training data quality. Lastly, we provide a qualitative analysis examining different categories of false negatives identified in training datasets.

Our contributions are as follows: (1) We are the first to report that carelessly adopting enormous training data may negatively affect the retriever and reranker model training. We show that the retrieval effectiveness can be improved by 4% with 57% *less* data, (2) We propose a LLM cascading framework that identifies and relabels the false hard negatives at an instance level. Our approach results in higher in-domain and out-of-domain retrieval effectiveness with a simpler training pipeline.

2 Related Work

Sparsely-annotated datasets. Popular IR training datasets, such as MS MARCO (Nguyen et al., 2016), were shallow pooled and sparsely judged by human assessors (Mackenzie et al., 2021; Arabzadeh et al., 2022). The assessor observed a few passages from a baseline retrieval system, picked those relevant to the query, and labeled them as ground-truth. On the other hand, non-relevant judged passages (i.e., passages seen but preferred lower than the ground truth) were not provided.

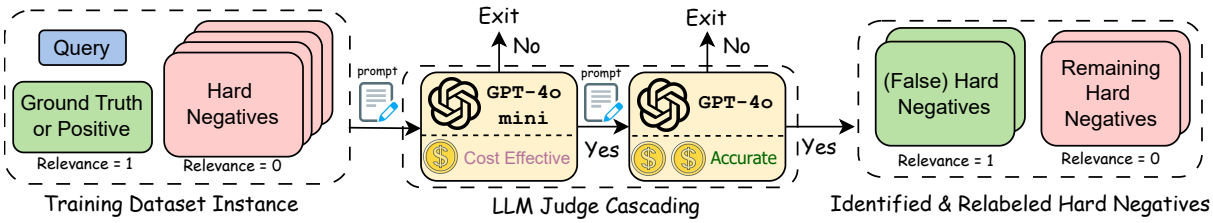

Figure 2: Flowchart for **RLHN** (**ReL**abeling **H**ard **N**egatives): (1) Provide the query, ground-truth or positive passages, and hard negative passages from a training instance as input, (2) Prompt a cost-effective LLM judge (e.g., GPT-4o-mini) and evaluate whether any hard negative is misclassified, (3) If yes, repeat the prompt with an accurate LLM judge (e.g., GPT-4o) (4) Output the relabeled hard negative passages (which are found relevant) and either remove them or relabel them as ground-truth passages in our experiments.

Therefore, an assumption is made in fine-tuning where remaining passages (in a passage corpus) are negatives, and a few mined passages similar to the query are labeled as hard negatives. In this work, we avoid relabeling false positives, as these labels are trustworthy, provided by a human assessor, who can have a different preference than the LLM itself.

LLM-based data curation. Hiring human assessors for judgments is expensive and time-consuming, and produces limited training pairs, e.g., 1K pairs in LIMA (Zhou et al., 2023). Alternatively, LLMs as judges have been recently explored for dataset curation in tasks, such as reranking (Ma et al., 2023; Zhuang et al., 2024; Qin et al., 2024), instruction fine-tuning (Chen et al., 2024b; Chen and Mueller, 2024), or even code-generation (Jain et al., 2024).

Pseudo-labeling. Instead of using supervised judgments, pseudo-labeling tackles the problem of sparse annotations by employing other techniques to estimate query-document relevance. Examples include distillation from cross-encoders (Qu et al., 2021), or ranking documents through prompting LLMs (Sun et al., 2023), or through composite measures of embedding similarity with ground-truth documents (Zerveas et al., 2023).

False negatives. Qu et al. (2021) first noted the issue of false negatives in retrieval, where certain hard negative passages should have been classified as positives. However, instead of *curating* the training datasets, RocketQA (Qu et al., 2021) fine-tuned models by distilling knowledge from the cross-encoder score for the query–document pair. Similarly, Moreira et al. (2024) examined various filtering methods for negative sampling by avoiding very hard negatives. In Gecko (Lee et al., 2024, 2025b), an LLM such as Gemini was used to relabel positive passages and identify better hard negatives. However, unlike our work, they focused on relabeling synthetic queries rather than existing collections like MS MARCO or NQ.

3 The RLHN Methodology

In this section, we discuss the LLM judge cascading framework, training dataset modifications, and dataset postprocessing and statistics.

3.1 LLM Judge Cascading Framework

We adopt a simple and cost-effective approach of using cascaded LLM judges (shown in Figure 2) inspired by Chen et al. (2024c) to identify false hard negatives datasets at a large scale. The framework involves two stages:

1. **Cost-effective judge (GPT-4o-mini):** In the first stage, we prompt GPT-4o-mini (OpenAI, 2024), a cost-effective LLM in the first stage to improve recall by identifying potential pairs with false negatives across all training pairs.
2. **Accurate judge (GPT-4o):** In the second stage, we prompt GPT-4o (OpenAI, 2024), a more reliable and expensive judge[3] to re-evaluate the potential pairs with false negatives identified by GPT-4o-mini and re-evaluate them using GPT-4o to improve precision.

3.2 Training Dataset Modification

Upon successful completion of identifying the false negatives, we compare three operations on the identified false negatives as follows:

- **Remove**: Discard the complete training instance due to the low quality, even if it contains at least one false negative.[4]

[3] GPT-4o-mini and GPT-4o pricing (as of May 15th, 2025) is 0.6$ and 5.0$ for 1M input tokens and 2.4$ and 20.0$ for 1M output tokens, respectively.

[4] We lose the instance completely in the "remove" technique.

Dataset	#Train Pairs	Avg. GT/Q	Avg. HN/Q	RLHN Stage 1	RLHN Stage 2
MS MARCO	485,823	1.1	25.0	391,965	326,301
HotpotQA	84,516	2.0	20.0	11,268	4,756
NQ	58,568	1.0	98.5	32,184	19,199
FEVER	29,096	1.3	20.0	7,764	3,577
SciDocsRR	12,655	1.6	19.7	2,068	351
FiQA-2018	5,500	2.6	15.0	3,632	1,833
ArguAna	4,065	1.0	13.6	0	0

Table 1: BGE training dataset statistics (Chen et al., 2024a). Avg. GT/Q denotes the average ground truth passages per query, and Avg. HN/Q denotes the average hard negative passages per query. RLHN Stages 1 & 2 show training pairs with at least one false hard negative.

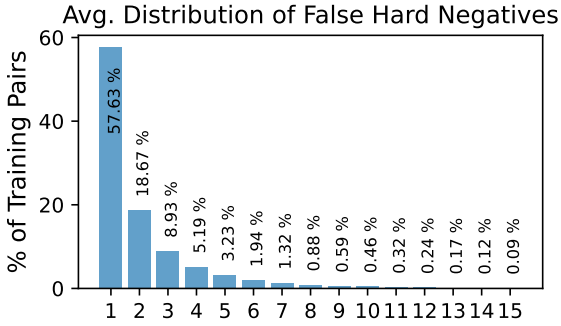

Figure 3: The distribution of training pairs (with at least one false negative) across false hard negatives detected. 58% of the training pairs detected contain a single false negative, 19% with two false negatives, and so on.

- **Remove HN**: Discard only the detected false negatives from the hard negative subset, keeping the instance with the remaining hard negatives.
- **Relabel HN (RLHN)**: Relabel only the detected false negatives from the hard negative subset, by adding them to the ground truth subset, keeping the instance with the remaining hard negatives.

3.3 Dataset Postprocessing & Statistics

In Table 1, we show the training dataset statistics observed in the BGE training collection. MS MARCO contains the highest amount of training pairs, followed by HotpotQA. All datasets contain training pairs with 1–3 ground-truth passages and 13–25 hard negatives (except NQ with 98–100 hard negatives).

False negatives. From Table 1, we see a majority of detected false negatives occur in MS MARCO (91.6% of all detected pairs). A maximum of up to 56% of all training pairs in MS MARCO contain false negatives, to a minimum of about 3% in SciDocsRR.[5] From Figure 3, we observe that in 58% of all detected false negative pairs, only a single false positive was detected, and 19% with two false negatives, and less than 1% with eight or more false negatives. If we detect any training pair with detected false negatives over a certain threshold k ($k = 7$ in our experiments), we excluded the pair completely in RLHN, as the query is likely to be ambiguous, that might not be a useful training instance (e.g., *what color is amber urine?*).

Cost estimates. We report the maximum costs incurred in RLHN (accurate input tokens + estimated 2048 output tokens on average) by both judges at each cascading stage: GPT-4o-mini and GPT-4o in Table 2. Overall, running RLHN with GPT-4o-mini in Stage 1 costs around $\approx$ 300 USD and with GPT-4o in Stage 2 costs around $\approx$ 3000 USD.

4 Experimental Setting

BGE training data. We utilize the original BGE training dataset[6] (Li et al., 2024), a comprehensive collection with training datasets for retrieval (e.g., NQ, MS MARCO), clustering (e.g., TwentyNewsgroups), and classification (e.g., AmazonReviews) tasks. Many of these training datasets are used in fine-tuning of popular retriever models such as E5-Mistral (Wang et al., 2024), GRIT-LM (Muennighoff et al., 2025), Linq (Choi et al., 2024), LLM2Vec (BehnamGhader et al., 2024), CDE (Morris and Rush, 2025), or NV-Embed (Lee et al., 2025a). Our work focuses on the *retrieval task*, therefore, we remove all training datasets from clustering and classification tasks, resulting in 15 datasets focused on the retrieval task, comprising a total of 1.6M training pairs, originally released with the MIT license.

LLM judges. In our work, we use GPT-4o-mini (version `2024-07-18`) and GPT-4o (version `2024-11-20`) as the judge using the Azure OpenAI service in the *batch* setting. We follow a temperature setting of 0.1 and use a chain-of-thought prompt setting (Wei et al., 2022). The prompt first evaluates the relevance between every hard negative passage and the question, and compares them with the ground truth to identify potential false negatives. We prompt up to 25 hard negative passages per query in a single API call as shown in Figure 6.

[5] We avoid relabeling ArguAna due to its inherent complex task, which doesn't measure directly for argument similarity, but rather counter arguments given an argument. Therefore, we keep the original dataset in fine-tuning without relabeling.

[6] huggingface.co/datasets/cfli/bge-full-data

	Cascading Stage 1		*Cascading Stage 2*	
Dataset	**# Pairs**	**GPT-4o-mini**	**# Pairs**	**GPT-4o**
MS MARCO	485,823	180.40 USD	391,965	2431.98 USD
HotpotQA	84,516	43.35 USD	11,268	97.26 USD
NQ	58,568	37.41 USD	32,184	345.08 USD
FEVER	29,096	22.67 USD	7,764	103.99 USD
SciDocsRR	12,655	9.07 USD	2,068	24.81 USD
FiQA-2018	5,500	3.60 USD	3,632	40.17 USD
Total Costs		**~300 USD**		**~3000 USD**

Table 2: Cost estimates for relabeling false negatives in RLHN using GPT-4o-mini and GPT-4o.

Evaluation benchmarks. We evaluate the retrieval and reranker accuracy of the models fine-tuned on datasets with false negatives either removed or relabeled with RLHN on the BEIR benchmark (Thakur et al., 2021) and AIR-Bench (Chen et al., 2025). Both benchmarks evaluate retrieval accuracy in nDCG@10. BEIR contains human-constructed datasets, and AIR-Bench contains datasets automatically generated by LLMs without human intervention. In BEIR, we drop Quora and CQADupstack and evaluate on the remaining 16 datasets. In AIR-Bench (version `24.05`), we evaluate five specific domains in English-only: Arxiv, Finance, Healthcare, Law, and News.

Backbone models. We use the E5 (base) unsupervised[7] (Wang et al., 2022b, 2024), a BERT-based encoder, due to its high accuracy on BEIR (preliminary results in Appendix A), the inclusion of a pre-training stage, and lower training complexity. E5 (base) contains 110M parameters, 12 layers, and a 768 embedding dimension with mean pooling. Also, we use a LLM-based decoder model with Qwen2.5-7B model[8] (Yang et al., 2024) with 7.61B parameters, 28 layers, and a 3584 embedding dimension with the `[EOS]` token pooling as the retrieval models. In addition, we use Qwen2.5-3B model (Yang et al., 2024)[9] for the reranker.

Fine-tuning details. All models were fine-tuned using 7 hard negatives, 1 positive, and random in-batch negatives (128 total) per batch, optimized with the InfoNCE loss function (van den Oord et al., 2018) using the Tevatron repository[10] (Gao et al., 2023; Ma et al., 2025) for up to 4–5 epochs, with a learning rate of 2e-5, and a maximum sequence length of 350 tokens (512 tokens during inference). We append a "`query: `" and "`passage: `" prefix. E5 (base) models are fine-tuned using 4×L40S GPUs, and Qwen2.5-7B and Qwen2.5-3B using a maximum of 2×H200 GPUs.

[7] `intfloat/e5-base-unsupervised` on HuggingFace.
[8] `Qwen/Qwen2.5-7B` on HuggingFace.
[9] `Qwen/Qwen2.5-3B` on HuggingFace.
[10] `https://github.com/texttron/tevatron`

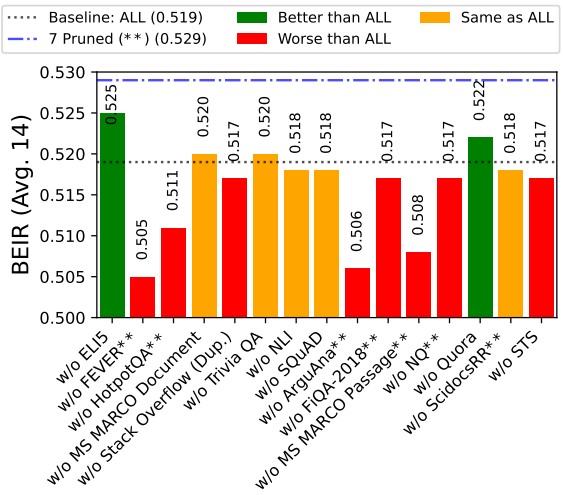

Figure 4: Dataset pruning by leaving one dataset out during fine-tuning E5 (base) on the BGE-training collection; [ALL] denotes fine-tuning on all datasets with 1.6M training pairs; [7 Pruned] denotes fine-tuning on 680K training pairs with seven remaining datasets (or 57.5% pairs) after dataset pruning. [Better than ALL] denotes the results *improved* after *removing* the dataset, meaning it has negative impact on the training process. [Worse than ALL] denotes the opposite, where the dataset has a positive impact on the training.

Baselines. To evaluate the impact of relabeling hard negatives using RLHN, we include two baselines: (1) *hard-negative mining*: Top-95% TopK-PecPos sampling (Moreira et al., 2024) on the default training dataset, using similarity scores computed for all hard negatives with the `bge-reranker-v2-gemma` reranker, and (2) *cross-encoder distillation*: we compute the normalized similarity scores for all query and hard negatives and positive pair on the default training dataset with the `bge-reranker-v2-gemma` reranker. We fine-tune the E5-base using knowledge distillation from the cross-encoder scores, with 1 positive, 15 hard and zero in-batch negatives using Tevatron.

5 Experimental Results

5.1 Preliminary Results: Dataset Pruning

False datapoint can hurt the training of retriever models. We assess the individual dataset contribution by evaluating several model variants by leaving one dataset out and fine-tuning the rest.

BEIR Dataset	*No Filtering*	*Baselines*		*Cascading Stage 1: GPT-4o-mini*			*Cascading Stage 2: GPT-4o-mini + GPT-4o*			*No Filtering*	*Cascading Stage 2*	
	Default	**TopK-PercPos**	**CE Distill**	**Remove**	**Remove HN**	**RLHN**	**Remove**	**Remove HN**	**RLHN**	**Default**	**Remove HN**	**RLHN**
Backbone	E5 (base)	E5 (base)	E5 (base)	E5 (base)	E5 (base)	E5 (base)	E5 (base)	E5 (base)	E5 (base)	Qwen2.5-7B	Qwen2.5-7B	Qwen2.5-7B
TREC-COVID†	0.783	0.789	0.793	0.786	0.793	0.798	0.794	0.785	0.809	0.797	0.771	0.815
NFCorpus†	0.378	0.377	0.363	0.378	0.380	0.381	0.380	0.382	0.390	0.389	0.389	0.391
NQ	0.595	0.601	0.624	0.593	0.592	0.602	0.573	0.598	0.591	0.597	0.602	0.623
HotpotQA	0.737	0.734	0.741	0.737	0.736	0.739	0.741	0.736	0.735	0.704	0.702	0.729
FiQA-2018	0.439	0.434	0.417	0.443	0.440	0.444	0.441	0.445	0.448	0.453	0.461	0.465
ArguAna	0.701	0.697	0.725	0.702	0.706	0.700	0.700	0.700	0.692	0.554	0.550	0.560
Touché-2020†	0.256	0.286	0.305	0.255	0.271	0.268	0.218	0.265	0.266	0.221	0.211	0.230
DBPedia	0.438	0.444	0.446	0.439	0.437	0.442	0.433	0.441	0.447	0.443	0.456	0.472
SCIDOCS	0.242	0.243	0.216	0.243	0.243	0.244	0.245	0.243	0.242	0.245	0.243	0.252
FEVER	0.878	0.878	0.889	0.875	0.876	0.877	0.881	0.876	0.871	0.863	0.857	0.872
Climate-FEVER	0.391	0.386	0.377	0.388	0.385	0.391	0.382	0.384	0.367	0.370	0.373	0.360
SciFact	0.735	0.735	0.727	0.741	0.731	0.733	0.744	0.735	0.740	0.755	0.755	0.767
TREC-NEWS†	0.465	0.466	0.458	0.470	0.466	0.473	0.464	0.473	0.484	0.494	0.480	0.487
Robust04†	0.442	0.451	0.452	0.448	0.452	0.471	0.447	0.458	0.497	0.501	0.501	0.540
Signal-1M (RT)†	0.275	0.272	0.271	0.279	0.275	0.275	0.274	0.270	0.274	0.275	0.268	0.280
BioASQ†	0.378	0.375	0.413	0.382	0.385	0.392	0.384	0.384	0.394	0.408	0.412	0.438
Avg. 16 (All)	0.508	0.511	0.514	0.510	0.511	0.514	0.506	0.511	**0.515**	0.504	0.502	**0.518**
Avg. 7 (OOD)	0.425	0.431	0.436	0.428	0.432	0.437	0.423	0.431	**0.445**	0.441	0.433	**0.454**

Table 3: Retrieval results measuring nDCG@10 on 16 datasets in the BEIR benchmark by fine-tuning retrieval models on variants of the BGE training dataset after relabeling false negatives. The seven unseen (or out-of-domain) datasets during fine-tuning are highlighted with † and their average scores are provided in Avg. 7.

As we fine-tune many models, i.e., one for each removed dataset, we limit these experiments to E5 (base). Summarized results are shown in Figure 4 (detailed results can be found in Table 12), demonstrating that training datasets (highlighted in red) can hurt the model retrieval accuracy, such as ELI5, removing which improves the nDCG@10 on BEIR (0.519 → 0.525). Also, it shows that certain datasets (highlighted in green) are crucial for model accuracy.

Based on findings in Figure 4 and selecting necessary datasets for individual task-based performances in BEIR, we prune the original 16 retrieval datasets in the BGE collection and select seven datasets (highlighted as $**$), reducing the training dataset size from 1.6M to 680K training pairs in our experiments. The average nDCG@10 score of E5 (base) improves from 0.519 → 0.529 on 14 datasets on average in BEIR, by fine-tuning on almost 2.35× smaller dataset (1.6M → 680K).

5.2 Main Results: Relabeling False Negatives

This section shows the results of the fine-tuned models on the variants of the training dataset described in Section 3.1 and 3.2, keeping the rest of the model training parameters unchanged.

BEIR benchmark. Results in Table 3 show that for both E5 (base) and Qwen2.5-7B, the RLHN technique achieves the best overall average nDCG@10 of 0.515 and 0.518 on 16 datasets on BEIR, outperforming models trained with the default setting and other remove techniques. The relabeled data in RLHN improves model generalization, with improvements strongly visible in seven out-of-domain (OOD) datasets in BEIR.

Backbone	Technique	Arxiv	Finance	Health.	Law	News	Avg. 5
E5 (base)	**Default**	0.345	0.401	0.521	0.117	0.455	0.368
E5 (base)	**TopK-PercPos**	0.348	0.418	0.529	0.119	0.464	0.376
E5 (base)	**CE Distill**	0.372	0.430	0.536	0.168	0.498	**0.401**
Cascading Stage 1: GPT-4o-mini							
E5 (base)	**Remove**	0.346	0.407	0.526	0.118	0.452	0.370
E5 (base)	**Remove HN**	0.344	0.406	0.522	0.118	0.459	0.370
E5 (base)	**RLHN**	0.362	0.421	0.522	0.123	0.465	**0.379**
Cascading Stage 2: GPT-4o-mini + GPT-4o							
E5 (base)	**Remove**	0.341	0.403	0.514	0.125	0.438	0.364
E5 (base)	**Remove HN**	0.346	0.411	0.525	0.124	0.464	0.374
E5 (base)	**RLHN**	0.356	0.440	0.521	0.138	0.476	**0.386**
Qwen2.5-7B	**Default**	0.325	0.391	0.479	0.115	0.430	0.348
Cascading Stage 2: GPT-4o-mini + GPT-4o							
Qwen2.5-7B	**Remove HN**	0.335	0.384	0.487	0.111	0.423	0.348
Qwen2.5-7B	**RLHN**	0.330	0.418	0.494	0.133	0.450	**0.365**

Table 4: Retrieval results measuring nDCG@10 on five specialized domains in AIR-BENCH dev (version `24.05`) by fine-tuning E5 (base) and Qwen2.5-7B on variants of the BGE training dataset with RLHN.

Stage 1 (RLHN) outperforms the Default setting by 2.0 points and Stage 2 (RLHN) by 3.2 points in nDCG@10. Overall, relabeling false negatives improves the data quality, which is reflected in model generalization across out-of-domain settings in BEIR.

AIR-BENCH. In addition to BEIR, AIR-BENCH provides a zero-shot setting to evaluate on challenging domains, such as Law. Table 4 shows the average nDCG@10 on five specialized domains. The improvements in model generalization are consistent to what we observed in BEIR. Stage 1 (RLHN) improves the Default setting by 1.1 points in nDCG@10, and Stage 2 (RLHN) further improves by 2.1 points. Overall, without changing the model or training parameters, mitigating false negatives in training datasets with RLHN enables the model generalize better to specialized domains in AIR-BENCH.

BEIR Dataset	*No Filtering*	*Cascading Stage 1*	*Cascading Stage 2*
	Default	**RLHN**	**RLHN**
TREC-COVID†	0.836	0.861	0.862
NFCorpus†	0.401	0.414	0.415
NQ	0.730	0.739	0.736
HotpotQA	0.863	0.861	0.861
FiQA-2018	0.517	0.521	0.519
ArguAna	0.740	0.730	0.763
Touché-2020†	0.275	0.308	0.313
DBPedia	0.532	0.536	0.538
SCIDOCS	0.278	0.273	0.270
FEVER	0.941	0.939	0.936
Climate-FEVER	0.457	0.468	0.430
SciFact	0.786	0.793	0.794
TREC-NEWS†	0.507	0.513	0.527
Robust04†	0.531	0.548	0.589
Signal-1M†	0.292	0.276	0.274
BioASQ†	0.510	0.505	0.500
Avg. 16 (All)	0.575	0.580	**0.583**
Avg. 7 (OOD)	0.479	0.489	**0.497**

Table 5: Reranker results measuring nDCG@10 on 16 datasets in BEIR by fine-tuning reranker models (based on Qwen2.5-3B) on variants of the BGE training datasets after relabeling false negatives. Stage 1 and 2 refers to GPT-4o-mini and GPT-4o-mini + GPT-4o.

Comparison with baselines. Results in Table 3 and Table 4 show that carefully avoiding sampling very hard negatives using Top-95%-PercPos outperforms the Default model, but still underperforms compared to the RLHN strategy with E5 (base). Next, the `bge-reranker-v2-gemma` cross-encoder, used as the distillation teacher is a strong baseline. It slightly underperforms RLHN on BEIR but outperforms RLHN on AIR-BENCH. However, we want to reiterate that our core motivation is to *identify* and *relabel* false negatives in training datasets to enhance data quality. Distillation-based fine-tuning requires on a strong, domain-focused cross-encoder reranker. Similarly, RLHN is particularly valuable for fine-tuning cross-encoders when teacher supervision is not viable.

Reranker results. Training data with improved quality also benefits cross-encoder rerankers. Table 5 shows the result comparison on the BEIR benchmark, where we rerank the top-100 results from the fine-tuned E5 (base) in the Default setting. Training rerankers with data fixed on RLHN Stages 1 and 2 progressively increases nDCG@10 on BEIR datasets by 0.5 points and 0.8 points, respectively. This improvement is most prominent on the seven OOD datasets, consistent with the above observation on retrievers: the data correction on the two stages improves the averaged OOD results by 1.0 and 1.8 points, respectively.

We note that the scale of the improvement on cross-encoders is not as large as on retrievers,

BEIR Dataset	*RLHN (Ablation of Hard Negatives)*			
	RLHN (1 HN)	**RLHN (3 HN)**	**RLHN (7 HN)**	**RLHN (9 HN)**
TREC-COVID†	0.809	0.810	0.809	0.812
NFCorpus†	0.389	0.388	0.390	0.392
NQ	0.563	0.583	0.591	0.595
HotpotQA	0.717	0.729	0.735	0.739
FiQA-2018	0.438	0.448	0.448	0.450
ArguAna	0.660	0.679	0.692	0.693
Touché-2020†	0.249	0.263	0.266	0.276
DBPedia	0.439	0.442	0.447	0.447
SCIDOCS	0.234	0.238	0.242	0.243
FEVER	0.851	0.864	0.871	0.875
Climate-FEVER	0.339	0.362	0.367	0.371
SciFact	0.736	0.737	0.740	0.744
TREC-NEWS†	0.481	0.473	0.484	0.484
Robust04†	0.506	0.502	0.497	0.499
Signal-1M†	0.273	0.272	0.274	0.272
BioASQ†	0.384	0.394	0.394	0.397
Avg. 16 (All)	0.504	0.512	0.515	**0.518**
Avg. 7 (OOD)	0.442	0.443	0.445	**0.447**

Table 6: Ablation of number of hard negatives during fine-tuning with InfoNCE loss function (van den Oord et al., 2018) in Tevatron with E5 (base).

which may indicate that cross-encoder rerankers are comparatively more robust to false negatives. However, albeit small, cross-encoders still benefit from training data of higher quality, especially when generalizing to unseen domains.

6 Analysis

Ablation on hard-negatives and significance tests. As an ablation, we experiment with the number of hard negatives during fine-tuning E5 (base) in Tevatron. From Table 6, we observe that increasing the number of hard negatives improves the nDCG@10 score on BEIR, with the best scores observed using 9 hard negatives.

We conduct statistical significance tests using ranger plots (Sertkan et al., 2023) for both E5 (base) and Qwen2.5-7B, comparing RLHN versus the Default setting. The ranger plots are provided in the Appendix (Figure 10 and Figure 11). In Figure 10, the plot shows statistical improvement for 10/16 BEIR datasets with E5 (base) fine-tuned using RLHN. Similarly, Figure 11 shows statistical improvement for 14/16 BEIR datasets with Qwen2.5-7B fine-tuned using RLHN.

Robustness of RLHN across varying training data subsets. As training datasets can be large, relabeling all training pairs using the LLM cascading pipeline can be computationally prohibitive. From Figure 5, we demonstrate that RLHN remains robust and maintains similar accuracy gains, even when applied to smaller randomly sampled subsets of the training dataset. To evaluate this, we use four random subsets (100K, 250K, 400K, and 680K) of the training datasets, with each dataset's distribution shown in Table 10.

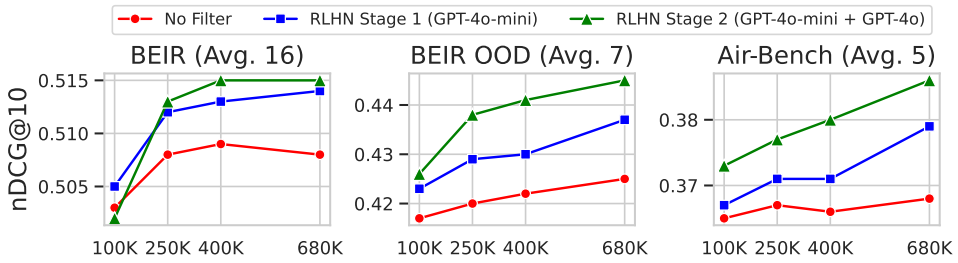

Figure 5: nDCG@10 scores on BEIR (Avg. 16 and Avg. 7) and AIR-BENCH (Avg. 5) by fine-tuning E5 (base) on a subset of the 100K, 250K, 400K, and 680K training pairs using the "RLHN" technique for both stages. All individual dataset scores for both BEIR and AIR-BENCH are provided in Figure 7 and Figure 8.

Datasets →	FEVER (3,521)		FIQA-2018 (1,829)		HOTPOTQA (4,720)		SCIDOCSRR (350)	
SoTA Reranker Judge ↓	**mAP@10**	**P@L(GT)**	**mAP@10**	**P@L(GT)**	**mAP@10**	**P@L(GT)**	**mAP@10**	**P@L(GT)**
`BAAI/bge-reranker-v2-gemma`	**0.839**	**0.777**	0.632	0.492	**0.742**	**0.638**	**0.926**	**0.875**
`mxbai/rerank-large-v2`	0.496	0.365	**0.658**	**0.525**	0.737	0.634	0.680	0.524
`mxbai/rerank-base-v2`	0.570	0.455	0.598	0.464	0.671	0.565	0.612	0.462
`Cohere (rerank-v3.5)`	0.811	0.740	0.572	0.437	0.694	0.588	0.838	0.743
`Alibaba-NLP/gte-reranker-modernbert-base`	0.688	0.602	0.545	0.408	0.658	0.560	0.843	0.754
`cross-encoder/ms-marco-MiniLM-L12-v2`	0.745	0.656	0.517	0.387	0.587	0.479	0.832	0.755

Table 7: Reranker as the judge as a baseline to identify RLHN false negatives in each training dataset (written along with the count of training pairs). **mAP@10** calculates the average precision of false negatives (labeled as positives) in the top-10 reranked results. **P@L(GT)** calculates the precision of false negatives present in top-k reranked results, where k varies in each query, measuring the count of false negatives detected using RLHN.

Overall, we have two main findings: (i) the E5 (base) model fine-tuned on RHLN Stages 1 and 2 training data, with false hard negatives relabeled as positives, *consistently* outperforms the Default setting, and (ii) the steeper slope in nDCG@10 demonstrates *continual improvement* across zero-shot domains, as the amount of training data increases, especially as observed in AIR-BENCH.

Reranker distillation is competitive but limited in detecting false negatives. A reranker, or cross-encoder, is commonly used in knowledge distillation to fine-tune a retriever model as an alternative to the traditional contrastive or InfoNCE loss function (Hofstätter et al., 2020; Qu et al., 2021; Wang et al., 2022a). This approach bypasses the original relevance judgments, relying instead on knowledge encoded within the reranker itself. Rather than using RLHN, we evaluate how well rerankers detect false negatives in training datasets. Specifically, we rerank the hard negatives for each training instance and compute two metrics: (i) mAP@10, which measures the average precision of false negatives in the top-10 results, and (ii) P@L(GT), which measures the precision of false negatives among the top-k results, where k equals the number of false negatives.

Table 7 reports results of six reranker judges from various sources across four datasets. We observe that the `bge-reranker-v2-gemma` judge achieves the highest scores amongst its counterparts in identifying false negatives labeled by RLHN (except on FIQA). However, on datasets such as FIQA-2018 and HOTPOTQA, rerankers detect only 52.5–63.8% of false negatives, indicating that while existing rerankers are competitive, they still require improvement. We suspect this limitation arises because rerankers are fine-tuned on these existing training datasets that contain false negatives, which negatively affects their accuracy.

Metric	**GPT-4o-mini**	**GPT-4o**
Cohen's Kappa (κ)	0.320	0.390

Table 8: Cohen's κ score of GPT-4o-mini and GPT-4o with human judgments on 670 query–negative pairs.

7 Human Validation

We conducted a validation study with three human assessors conducting using Label Studio[11] for data annotation. The assessors were briefed on the relevance task, and then independently evaluated a total of 670 query–hard negative pairs. The hard negatives were randomly sampled from the RLHN set, each containing at least one false negative. During the assessment, all annotators worked independently and were not exposed to the LLM

[11] `github.com/HumanSignal/label-studio`

Query	Ground Truth or Positive Passages		False Negatives (Detected by RLHN)	
(Q1) Which is a food magazine, Latin Mass Magazine or Saveur?	**Latin Mass Magazine**: A Journal of Catholic Culture, commonly referred to as Latin Mass Magazine, is an American Catholic magazine published quarterly, with a traditionalist Catholic viewpoint. [...]	**Saveur**: Saveur is a gourmet, food, wine, and travel magazine that specializes in essays about various world cuisines. Its slogan—"Savor a World of Authentic Cuisine"—signals the publication's focus on enduring culinary traditions [...]	**Food & Wine**: Food & Wine is a monthly magazine published by Time Inc. It was founded in 1978 by Ariane and Michael Batterberry. It features recipes, cooking tips, travel information, restaurant reviews, chefs, wine pairings and seasonal content [...]	**Cocina (magazine)**: is a Colombian-based monthly magazine published by Publicaciones Semana S.A.. It features recipes, cooking tips, culinary tourism information, restaurant reviews, chefs, wine pairings and seasonal holiday content [...]
(Q2) What year was the premier professional ice hockey league in the world established?	**2016–17 Minnesota Wild season**: The 2016–17 Minnesota Wild season was the 17th season for the National Hockey League franchise that was established on June 25, 1997.	**National Hockey League**: The National Hockey League (NHL; French: "Ligue nationale de hockey—LNH") is a professional ice hockey league currently comprising 31 teams [...]	**History of the National Hockey League (1917–42)**: History of the National Hockey League (1917–42) The National Hockey League (NHL) was founded in 1917 following the demise of its predecessor league, the National Hockey Association (NHA). [...]	
(Q3) name meaning yin and yang	**Yin and yang**: In Chinese philosophy, yin and yang (also, yin-yang or yin yang) describes how apparently opposite or contrary forces are actually complementary, interconnected, and interdependent in the natural world, and how they give rise to each other as they interrelate to one another.		**Yin and yang**: Yin and Yang are ancient Chinese philosophical terms, with the Yin Yang Theory being a fundamental part of Feng Shui. It is a Chinese theory on the perspective of continuous change and balance. [...]	**Yin Yang Symbols and Their Meanings**: In a nutshell, Chinese yin yang symbols represent perfect balance. A great deal of Chinese philosophy stems from the concept of yin and yang - opposites interacting [...]
(Q4) Charles, Prince of Wales is patron of numerous other organizations.	**Charles, Prince of Wales**: Charles, Prince of Wales (born 14 November 1948) is the eldest child and heir apparent of Queen Elizabeth II [...] Charles's interests encompass a range of humanitarian and social issues: he founded The Prince's Trust in 1976, sponsors The Prince's Charities, and is patron of numerous other charitable and arts organisations. [...]		**Julia Cleverdon Dame**: Julia Charity Cleverdon [...] served for 16 years as Chief Executive of Business in the Community, one of the Prince's Charities of Charles, Prince of Wales.	**The Prince's Trust**: The Prince's Trust is a charity in the United Kingdom founded in 1976 by Charles, Prince of Wales, and Frederick John Pervin to help young people. [...]

Table 9: Qualitative analysis showcasing the different varieties of false negatives detected by RLHN. The first two questions are taken from HOTPOTQA, the third from MS MARCO, and the last from FEVER. The text supporting the query is highlighted in green, partially supporting in orange, and not supporting with red.

predictions. An example of the annotation interface is shown in Figure 9.

Table 8 reports Cohen's Kappa (κ) measuring agreement between each LLM's predictions and the human labels. The κ scores are consistent with prior work reporting similar levels of human–LLM agreement (Arabzadeh and Clarke, 2025). GPT-4o shows substantially higher agreement with human annotators compated to GPT-4o-mini. This finding aligns with our empirical results, where relabeling with GPT-4o shows consistent gains over GPT-4o-mini in training retrieval and reranker models.

8 Qualitative Analysis of False Negatives

We qualitatively analyze the labeling accuracy of our LLM cascading framework by manually spot-checking a few training instances. As shown in Table 9, we observe a variety of false negatives, which fall into the following scenarios:

1. Detected false negatives are incorrect or not relevant. GPT-4o can sometimes detect a false negative that is not relevant to the query. E.g., (Q1) query asks which is a food magazine between *Latin Mass* or *Saveur*, however, the detected false negatives identify different food magazines such as *Food & Wine* or *Cochina*, which are both incorrect.

2. The ground truth may be incorrectly labeled. In a few queries, we observe that the ground truth passage can contain conflicting information with the false negative, resulting in incorrect labeling. E.g., the correct answer to the (Q2) query, which asks about the professional ice hockey establishment is 1917 (present in the false negative). However, the ground truth incorrectly states 1997.

3. The query may be too generic or ambiguous. In a substantial amount of training pairs in MS MARCO, we find that the training query is rather ambiguous, leading to many false negatives being detected. E.g., for the (Q3) query, all passages–including both the ground truth and false negatives–are relevant, as they each correctly define "yin and yang" but with different interpretations.

4. False negatives can be partially correct. Not all detected false negatives are entirely non-relevant to the query. E.g., one false negative is partially relevant to (Q4), which asks about organizations associated with Charles, the Prince of Wales.

9 Conclusion

In this work, we emphasize the importance of clean training datasets. First, we showed that certain datasets can negatively impact model effectiveness when fine-tuned across a huge collection with many training pairs. Dataset pruning removes 57.5% (8 datasets out of 15) and improves the model accuracy on BEIR by even 1.0 point and making the dataset 2.35× smaller. Next, after pruning, we observed the issue of false hard negatives in the remaining training datasets, where passages in the hard negative list are misclassified and are relevant to the query. We presented RLHN, an effective cascading LLM approach for relabeling hard negatives as ground truth or positives.

Using RLHN, both retrievers and rerankers consistently improved their model generalization on BEIR and zero-shot AIR-BENCH evaluations, as supported by human annotation results, outperforming competitive baselines such as hard-negative sampling and cross-encoder knowledge distillation.

Limitations

Even though we propose an effective technique to identify and relabel false hard negatives with RLHN, no technique is perfect and has its limitations. Making those explicit is a critical point in understanding the RLHN results and improvements, and for future work, to propose even better detection techniques.

1. False positives in training datasets. Detecting and relabeling false positives in training datasets is an important avenue of potential research. However, we avoid checking for false positives, as these labels are trustworthy, provided by a human assessor, who can have a different preference than the LLM itself. False positives might occur in a dataset due to human errors in existing datasets, but we suspect both the importance and frequency of detected false positives to be much lower than false negatives.

2. Cleaning extremely large training datasets. The maximum training dataset size that we covered in our work contained $\leq$1M training pairs. This is a reasonable dataset size to apply RLHN within a strict compute budget. Cleaning extremely large training datasets (for example, containing between 1–10M training pairs) is not feasible, as it may require a very high computation budget, with detection using GPT-4o. In the future, we wish to experiment with open-source LLMs, such as Qwen-3 (Yang et al., 2025), as an alternative in our LLM cascading pipeline, allowing relabeling of extremely large training datasets.

3. Multilingual and long-context document retrieval datasets. A majority of the training datasets included in the BGE training collection have average document lengths up to a few hundred words, roughly equivalent to a few paragraphs. Applying RLHN to clean long-context document retrieval datasets, such as MLDR (Chen et al., 2024a) and multilingual training datasets, such as MIRACL (Zhang et al., 2023), would be highly relevant in the future.

4. Multi-vector retrieval models. A popular suite of retrieval models includes multi-vector models, such as ColBERT (Khattab and Zaharia, 2020; Santhanam et al., 2022), representing queries and documents by multiple contextualized token-level embeddings. In our work, we limited our experiments to dense retrievers and rerankers. We keep RLHN with an extension to multi-vector models as future work, using a training repository such as PyLate (Chaffin and Sourty, 2025).

Acknowledgments

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada. Additional funding is provided by Microsoft via the Accelerating Foundation Models Research program.

References

Loubna Ben allal, Anton Lozhkov, Elie Bakouch, Gabriel Martin Blazquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Agustín Piqueres Lajarín, Hynek Kydlíček, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan Son NGUYEN, Ben Burtenshaw, Clémentine Fourrier, Haojun Zhao, Hugo Larcher, Mathieu Morlon, Cyril Zakka, and 3 others. 2025. SmolLM2: When Smol goes big — data-centric training of a fully open small language model. In *Second Conference on Language Modeling*.

Negar Arabzadeh and Charles L. A. Clarke. 2025. A human-AI comparative analysis of prompt sensitivity in LLM-based relevance judgment. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*, pages 2784–2788. ACM.

Negar Arabzadeh, Alexandra Vtyurina, Xinyi Yan, and Charles L. A. Clarke. 2022. Shallow pooling for sparse labels. *Inf. Retr. J.*, 25(4):365–385.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Antoine Chaffin and Raphaël Sourty. 2025. Pylate: Flexible training and retrieval for late interaction models. *CoRR*, abs/2508.03555.

Jianlyu Chen, Nan Wang, Chaofan Li, Bo Wang, Shitao Xiao, Han Xiao, Hao Liao, Defu Lian, and Zheng Liu. 2025. AIR-Bench: Automated heterogeneous information retrieval benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19991–20022, Vienna, Austria. Association for Computational Linguistics.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*,

pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Jiuhai Chen and Jonas Mueller. 2024. Automated data curation for robust language model fine-tuning. *CoRR*, abs/2403.12776.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2024b. AlpaGasus: Training a better alpaca with fewer data. In *The Twelfth International Conference on Learning Representations*.

Lingjiao Chen, Matei Zaharia, and James Zou. 2024c. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*.

Chanyeol Choi, Junseong Kim, Seolhwa Lee, Jihoon Kwon, Sangmo Gu, Yejin Kim, Minkyung Cho, and Jy-yong Sohn. 2024. Linq-Embed-Mistral technical report. *CoRR*, abs/2412.03223.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Tevatron: An efficient and flexible toolkit for neural retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 3120–3124, New York, NY, USA. Association for Computing Machinery.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2020. Improving efficient neural ranking models with cross-architecture knowledge distillation. *CoRR*, abs/2010.02666.

Naman Jain, Tianjun Zhang, Wei-Lin Chiang, Joseph E. Gonzalez, Koushik Sen, and Ion Stoica. 2024. LLM-assisted code cleaning for training accurate code generators. In *The Twelfth International Conference on Learning Representations*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

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, page 39–48, New York, NY, USA. Association for Computing Machinery.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025a. NV-Embed: Improved techniques for training LLMs as generalist embedding models. In *The Thirteenth International Conference on Learning Representations*.

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. 2025b. Gemini embedding: Generalizable embeddings from Gemini. *CoRR*, abs/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 Hernández Ábrego, 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. *CoRR*, abs/2403.20327.

Chaofan Li, Minghao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Yingxia Shao, Defu Lian, and Zheng Liu. 2024. Making text embedders few-shot learners. *CoRR*, abs/2409.15700.

Xueguang Ma, Luyu Gao, Shengyao Zhuang, Jiaqi Samantha Zhan, Jamie Callan, and Jimmy Lin. 2025. Tevatron 2.0: Unified document retrieval toolkit across scale, language, and modality. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 4061–4065, New York, NY, USA. Association for Computing Machinery.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-shot listwise document reranking with a large language model. *CoRR*, abs/2305.02156.

Joel Mackenzie, Matthias Petri, and Alistair Moffat. 2021. A sensitivity analysis of the MSMARCO passage collection. *CoRR*, abs/2112.03396.

Gabriel Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. 2024. NV-Retriever: Improving text embedding models with effective hard-negative mining. *CoRR*, abs/2407.15831.

John X Morris and Alexander M Rush. 2025. Contextual document embeddings. In *The Thirteenth International Conference on Learning Representations*.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2025. Generative representational instruction tuning. In *The Thirteenth International Conference on Learning Representations*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

OpenAI. 2024. Hello GPT-4o.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518, Mexico City, Mexico. Association for Computational Linguistics.

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. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847, Online. Association for Computational Linguistics.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3715–3734. Association for Computational Linguistics.

Mete Sertkan, Sophia Althammer, and Sebastian Hofstätter. 2023. Ranger: A toolkit for effect-size based multi-task evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 581–587, Toronto, Canada. Association for Computational Linguistics.

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

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, Singapore. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

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

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022b. Text embeddings by weakly-supervised contrastive pre-training. *CoRR*, abs/2212.03533.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11897–11916. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le,

and Denny Zhou. 2022. Chain-of-Thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

George Zerveas, Navid Rekabsaz, and Carsten Eickhoff. 2023. Enhancing the ranking context of dense retrieval through reciprocal nearest neighbors. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10779–10803, Singapore. Association for Computational Linguistics.

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

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

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems*, volume 36, pages 55006–55021. Curran Associates, Inc.

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. 2024. A setwise approach for effective and highly efficient zero-shot ranking with large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 38–47, New York, NY, USA. Association for Computing Machinery.

A Pretrained or Backbone Choice

We experimented with several pretrained or base model choices. In particular, we focused on fine-tuning recently introduced encoder models such as ModernBERT (Warner et al., 2025) to decoder-based large language models such as Qwen-2.5 (less than <500M parameters). We fine-tune each backbone on the whole BGE retrieval training subset (15 datasets & 1.6M training pairs) for up to 3 training epochs with different hyperparameters to fit the training with 4×A6000 GPUs. We plot the model configurations and training settings in Table 11.

Validation results. From Table 11, we observe that encoder models pre-trained such as E5-base or E5-large achieve the highest nDCG@10 scores on four BEIR datasets. These outperform recent backbones such as ModernBERT-base (Warner et al., 2025) or even smaller-sized LLMs such as Qwen-2.5 (0.5B). This anecdotally confirms that the unsupervised pre-training stage in E5 pretrained models is useful and necessary for achieving a competitive nDCG@10 score on BEIR. Since fine-tuning E5 (large) is around 2× slower than fine-tuning E5 (base), we run our main experiments on E5 (base) due to computational budget constraints.

Dataset	∼100K	∼250K	∼400K	∼680K
MS MARCO	49,571	145,000	210,000	485823
HotpotQA	10,250	30,000	84,516	84516
NQ	6110	30,000	58,568	58,568
FEVER	8017	28,755	28,755	28,755
SciDocsRR	12,654	12,654	12,654	12,654
FiQA	5500	5,500	5,500	5,500
ArguAna	4065	4,065	4,065	4,065
Total Pairs	**96,167**	**255,974**	**404,058**	**679,881**

Table 10: Training pair distribution across seven datasets for four configurations: 100K, 250K, 400K, and 680K.

B Leave-One-Dataset-Out Results

We provide detailed scores for leave-one-dataset-out (Figure 4) in Table 12, where we fine-tune E5-base retriever models on:

Part (a): no datasets;
Part (b): all 15 datasets;
Part (c): all 15 datasets but one left-out dataset;
Part (d): 7 datasets with the most significant effectiveness drop after being removed;

Backbone	#Params	#Layers	Hidden Size	Pool	LR	Batch Size	Epoch	Time Taken	COVID	NFC.	FiQA	SciFact
E5-large (unsup.) (Wang et al., 2022b)	330M	24	1024	*mean*	$1e-5$	128 x 8 x 4	3	$\sim$ 36 hours	<u>0.712</u>	**0.383**	**0.475**	**0.747**
ModernBERT-base (Warner et al., 2025)	149M	22	768	*mean*	$2e-4$	256 x 8 x 4	3	$\sim$ 12 hours	0.560	0.279	0.440	0.602
E5-base (unsup.) (Wang et al., 2022b)	110M	12	768	*mean*	$2e-5$	256 x 8 x 4	3	$\sim$ 18 hours	**0.731**	<u>0.381</u>	<u>0.444</u>	<u>0.728</u>
E5-small (unsup.) (Wang et al., 2022b)	33M	12	384	*mean*	$3e-5$	256 x 8 x 4	3	$\sim$ 13 hours	0.667	0.349	0.420	0.698
Qwen-2.5-0.5B (Yang et al., 2024)	500M	24	896	*last*	$1e-5$	96 x 8 x 4	3	$\sim$ 36 hours	0.503	0.356	0.417	0.692
SmolLM2-360M (allal et al., 2025)	360M	32	960	*last*	$1e-5$	96 x 8 x 4	3	$\sim$ 33 hours	0.670	0.336	0.355	0.635
SmolLM2-135M (allal et al., 2025)	135M	30	576	*last*	$1e-5$	128 x 8 x 4	3	$\sim$ 24 hours	0.668	0.327	0.304	0.608

Table 11: Model configuration, training settings, and retrieval results (nDCG@10) for backbone models fine-tuned on the BGE-training dataset (1.6M training pairs) and evaluated on four datasets from the BEIR benchmark. The models are sorted according to parameter size; The best score is highlighted as **bold**, the second best is <u>underlined</u>. COVID denotes the TREC-COVID dataset and NFC. denotes the NFCorpus dataset.

	Setting	Training Pairs	TREC-COVID	NFCorpus	NQ	HotpotQA	FiQA-2018	ArguAna	Touché-2020	DBPedia	SCIDOCS	FEVER	Climate-FEVER	SciFact	TREC-NEWS	Robust04	Avg. 14	Improved	Reduced	Keep Dataset?
(a)	**Pre-trained (Only)**	0	0.610	0.358	0.390	0.524	0.401	0.422	0.169	0.354	0.211	0.634	0.154	0.737	0.441	0.416	0.416	-	-	-
(b)	**(ALL) Training Pairs**	1.60M	0.731	0.381	0.595	0.726	0.444	0.652	0.181	0.437	0.233	0.871	0.370	0.728	0.434	0.477	0.519	-	-	-
	w/o ELI5	1.27M	0.772	0.378	0.593	0.728	0.424	0.652	0.213	0.434	0.235	0.868	0.377	0.734	0.469	0.478	0.525	7	5	✗
	w/o FEVER	1.57M	0.748	0.379	0.598	0.725	0.446	0.647	0.175	0.434	0.234	0.787	0.240	0.749	0.423	0.483	0.505	6	5	✓
	w/o HotpotQA	1.51M	0.724	0.381	0.600	0.642	0.449	0.652	0.178	0.425	0.232	0.863	0.358	0.725	0.441	0.489	0.511	4	7	✓
	w/o MS MARCO Document	1.23M	0.742	0.380	0.586	0.726	0.445	0.656	0.175	0.435	0.235	0.866	0.347	0.742	0.458	0.490	0.520	6	5	✗
	w/o Stack Overflow (Dup.)	1.58M	0.720	0.379	0.593	0.726	0.444	0.650	0.174	0.436	0.235	0.870	0.368	0.729	0.431	0.487	0.517	7	2	✗
	w/o Trivia QA	1.54M	0.729	0.380	0.595	0.730	0.450	0.647	0.174	0.440	0.234	0.870	0.382	0.731	0.443	0.481	0.520	7	3	✗
	w/o NLI	1.60M	0.729	0.380	0.594	0.726	0.445	0.652	0.177	0.437	0.233	0.870	0.368	0.728	0.436	0.477	0.518	1	3	✗
(c)	**w/o SQuAD**	1.51M	0.709	0.379	0.598	0.723	0.445	0.654	0.181	0.437	0.234	0.872	0.376	0.729	0.439	0.481	0.518	5	3	✗
	w/o ArguAna	1.59M	0.736	0.381	0.598	0.728	0.448	0.434	0.174	0.434	0.234	0.871	0.378	0.731	0.445	0.486	0.506	8	3	✓
	w/o FIQA-2018	1.59M	0.728	0.380	0.596	0.727	0.428	0.658	0.174	0.436	0.235	0.871	0.370	0.729	0.433	0.477	0.517	3	2	✓
	w/o MS MARCO Passage	1.11M	0.699	0.377	0.551	0.730	0.440	0.650	0.162	0.407	0.237	0.869	0.338	0.733	0.431	0.484	0.508	3	10	✓
	w/o NQ	1.54M	0.745	0.381	0.553	0.728	0.451	0.659	0.178	0.435	0.234	0.867	0.369	0.728	0.435	0.472	0.517	5	4	✓
	w/o Quora	1.54M	0.759	0.382	0.599	0.727	0.451	0.653	0.185	0.436	0.234	0.867	0.371	0.729	0.436	0.481	0.522	6	1	✗
	w/o SciDocsRR	1.59M	0.733	0.378	0.595	0.727	0.447	0.662	0.178	0.436	0.201	0.868	0.374	0.740	0.434	0.475	0.518	5	4	✓
	w/o STS	1.60M	0.718	0.379	0.596	0.727	0.446	0.652	0.177	0.437	0.234	0.867	0.369	0.729	0.435	0.478	0.517	1	4	✗
(d)	**7 Datasets Pruned (✓)**	680K	0.781	0.376	0.593	0.728	0.421	0.664	0.242	0.440	0.204	0.875	0.397	0.748	0.467	0.464	0.529	9	5	-

Table 12: Retrieval results measuring nDCG@10 on 14 datasets in the BEIR benchmark by fine-tuning E5 (base) by **leaving out one training dataset at a time** and fine-tuning the rest. Improved denotes E5 (base) with a nDCG@10 better than $+1$ point, Reduced with a nDCG@10 worse than -1 point, and No Change within the ± 1 point range, compared to part (b) E5 (base) fine-tuned on ALL Training Pairs. Each row in part (c) is fine-tuned on all but one left-out dataset. Part (c) is fine-tuned on the 7 selected datasets.

SYSTEM: Given (1) a search question, (2) a relevant ground-truth document, (3) and a set of unrelated documents that may appear in any system's response to that question. Your task is to evaluate whether any of the unrelated documents are relevant compared to the ground-truth document in answering the question. A document is only considered **relevant** to the question if it provides sufficient information in answering the question.

Input

You will receive:
1. *question*: The question that the to-be-judged documents will be evaluated on.
2. *ground_truth*: A pre-validated document judged as **most relevant** to the question. This document can answer the question and should be used as a guide for your analysis.
3. *documents*: A set of unrelated documents which may not be relevant in answering the question.

You will first read the question and carefully analyze each unrelated documents provided to you.
Read every question and unrelated document carefully as you would when proofreading.

Criteria

Use the following criteria to judge the relevance of each document:
- *Relevant*: A document is considered **relevant** to the question if it provides sufficient information in answering the question, containing **all** necessary parts highlighted in the ground truth.
- *Not Relevant*: The document does not answer the question and **does not** provide information in entailing parts present in the ground truth.

Output

Follow these detailed steps and output your reasoning for each step wrapped for each respective XML tag below:
1. You should think and provide your reasoning under <thinking> [. . .] </thinking> on **why** and **how** if an unrelated document is **relevant** following the criteria above.
2. Next, for all unrelated documents which are found to be **relevant**, compare them against the ground truth (<ground_truth>) document in answering the question under <preference> [. . .] </preference> tokens.
3. Finally, output the list of documents which are (1) relevant and (2) prefer better or equal under the XML tag (<better>) or worse (<worse>) than the ground truth (<ground_truth>) document for answering the question in <verdict> [. . .] </verdict>. Output [ ] if none of the documents are found to be relevant.

Follow strictly the format below:
<thinking> Evaluate the reasoning individually for all unrelated documents to answer the question
Doc (1): output the reasoning here
Doc (2): output the reasoning here
. . .
</thinking>
<preference> Compare the ground truth and every **relevant** document individually to answer the question
Doc (1): compare the relevance of Doc (1) with the <ground_truth> document here, which is more preferred?
. . .
</preference>
<verdict>
<better> Preferred over or equally as ground truth: [Doc (2) . . .] </better>,
<worse> Relevant but not preferred over ground truth: [Doc (1) . . .] </worse>
</verdict>

———

<question> {$question$} </question>
<ground_truth> {$ground\_truth$} </ground_truth>
<documents> {$documents$} </documents>

Figure 6: Prompt used in RLHN with GPT-4o-mini and GPT-4o for relabeling hard negatives for all BGE training datasets. Certain texts above in the prompt are bolded and tab-aligned to assist with reading. For both GPT-4o-mini (stage 1) and GPT-4o (stage 2) experiments, we consider negatives present within the <better> and </better> tags as false negatives. However, a training instance with any hard negative in either <better> and </better> or <worse> and </worse> tags in the first stage output (GPT-4o-mini judge) was forwarded to the second stage (GPT-4o judge) in the RLHN framework.

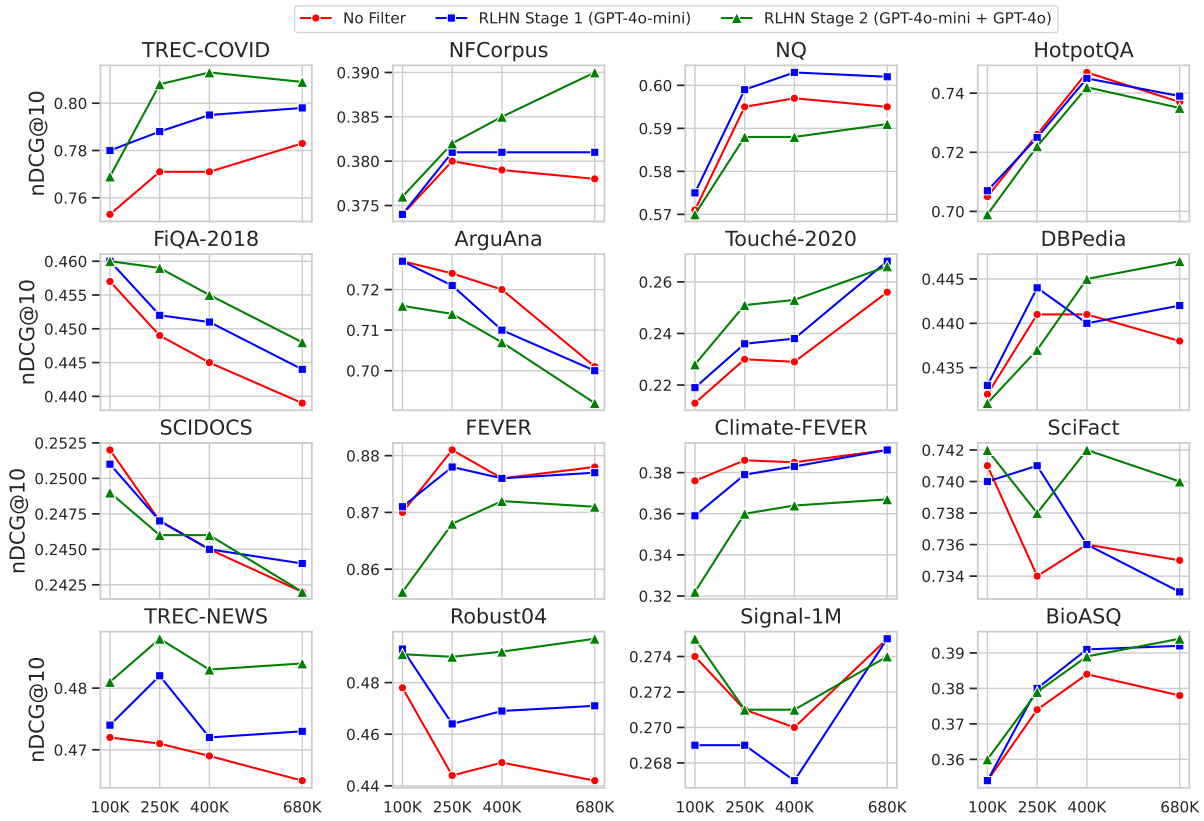

Figure 7: nDCG@10 scores on all 16 BEIR datasets by fine-tuning E5 (base) retrieval model on a subset of the 100K, 250K, 400K, and 680K training pairs from both stages 1 and 2, sampled from seven datasets in the BGE collection (listed in Table 1) using the RLHN framework. The training pair distribution is shown in Table 10.

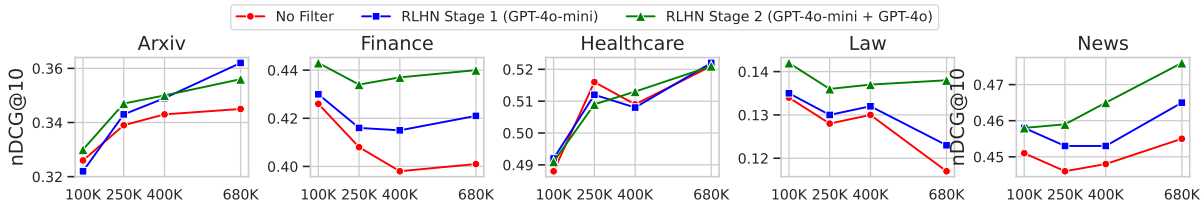

Figure 8: nDCG@10 scores on all 5 AIR-BENCH datasets by fine-tuning E5 (base) retrieval model on a subset of the 100K, 250K, 400K, and 680K training pairs from both stages 1 and 2, sampled from seven datasets in the BGE collection (listed in Table 1) using the RLHN framework. The training pair distribution is shown in Table 10.

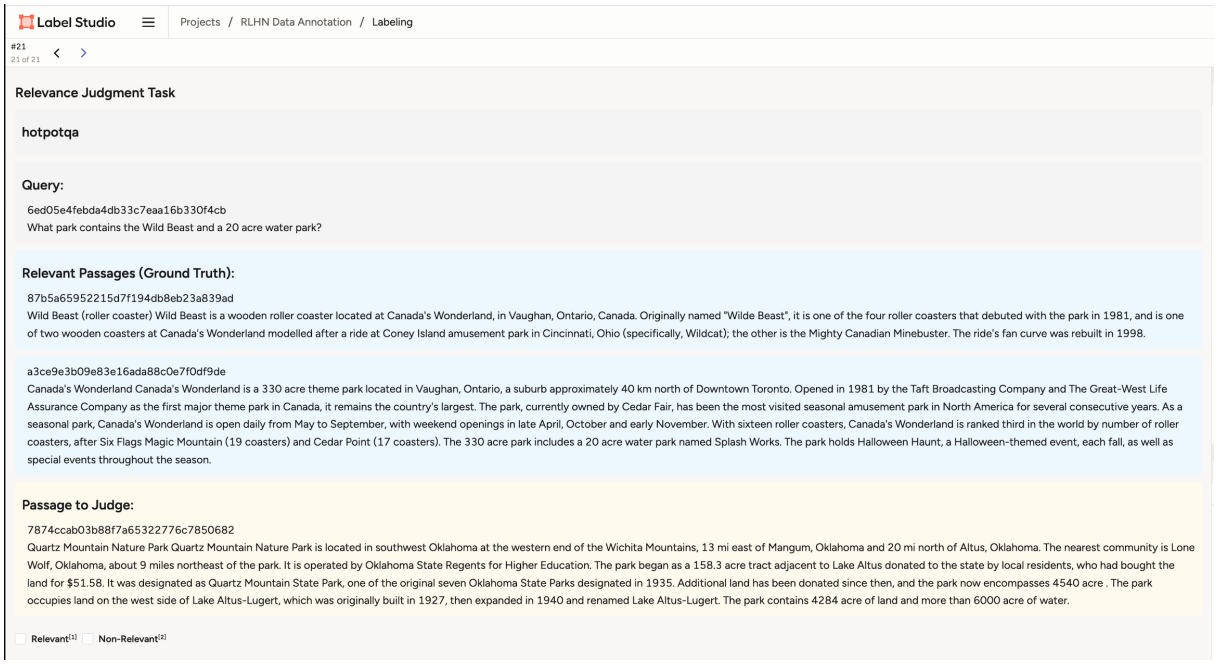

Figure 9: A screenshot of the human validation study conducted via Label Studio. First, the human assessor reads the query (highlighted in grey) and the relevant passages (highlighted in blue). Next, the assessor reads a sequence of hard negative passages one by one (highlighted in yellow) and evaluates the relevancy with the question, marking their decision in the checkbox as either (1) *relevant* or (2) *non-relevant*.

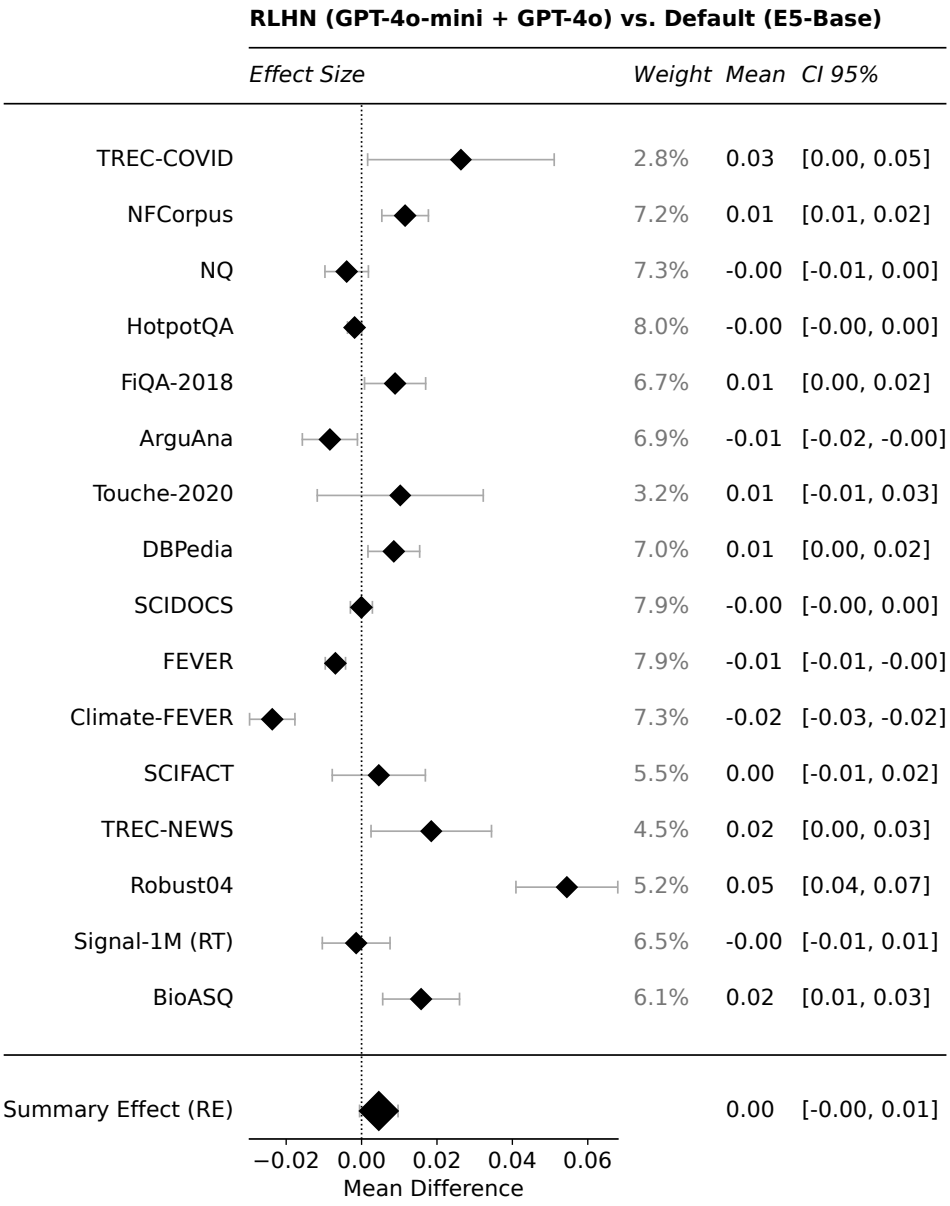

Figure 10: Ranger plot (Sertkan et al., 2023) showing the statistical significance of improvements observed in RLHN (GPT-4o-mini + GPT-4o) versus the Default setting for the E5 (base) fine-tuned retriever.

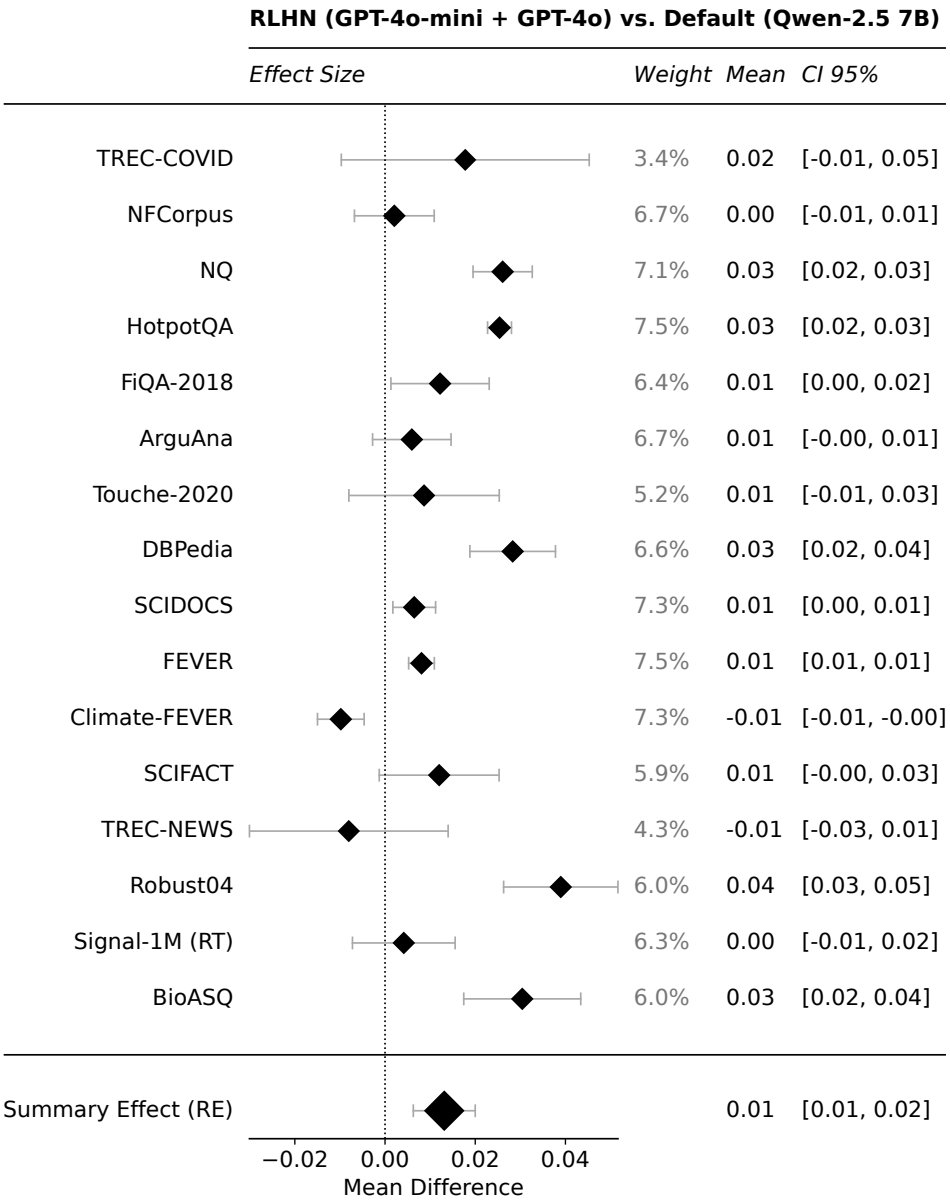

Figure 11: Ranger plot (Sertkan et al., 2023) showing the statistical significance of improvements observed in RLHN (GPT-4o-mini + GPT-4o) versus the Default setting for the Qwen2.5-7B fine-tuned retriever.