CLEAR: Consistent Labeling Enhanced by LLM-driven Automated Re-labeling for Improved Information Retrieval

Anonymous ACL submission

Abstract

The performance of information retrieval (IR) systems is heavily influenced by the quality of training data. Manually labeled datasets often contain errors due to subjective biases of annotators, and limitations of retrieval models. To address these challenges, we propose CLEAR, a novel framework that leverages large language models (LLMs) to automatically correct incorrect labels and extract more accurate and true positive documents. CLEAR estimates the reliability of existing annotations using LLMs and rectifies potential labeling errors, thereby improving overall data quality. Furthermore, we conduct a systematic investigation of how utilizing true positive documents affects retrieval model performance. We evaluate CLEAR on several widely-used IR benchmarks, including MS MARCO Passage, MS MARCO Document, Natural Questions, and TriviaQA. Experimental results demonstrate that CLEAR consistently outperforms existing baseline models, validating the effectiveness of the proposed approach.

1 Introduction

Natural language processing (NLP) tasks, such as question answering (QA) and information retrieval (IR), typically rely on manually annotated datasets. However, the manual annotation process is inherently susceptible to labeling errors and noise, arising from various factors such as annotator subjectivity, ambiguous annotation guidelines, cognitive biases, and occasional lapses in attention (Northcutt et al., 2021; Sheng et al., 2008; Snow et al., 2008; Paullada et al., 2021).

The issue becomes even more pronounced in crowd-sourced annotations involving non-expert workers, where label noise and inconsistencies are substantially more prevalent compared to expert-generated annotations (Zhang et al., 2025; Jamison and Gurevych, 2015). In tasks such as information retrieval (IR), which require relevance judgments, crowd workers often apply divergent criteria, leading to highly inconsistent labeling (Guo et al., 2023). Numerous studies have demonstrated that crowd-sourced annotations are significantly noisier than those produced by trained assessors (Chong et al., 2022). Furthermore, several widely used benchmark datasets have been shown to contain a non-negligible number of incorrect labels. Therefore, enhancing dataset quality is essential for the development of robust and reliable natural language processing (NLP) and information

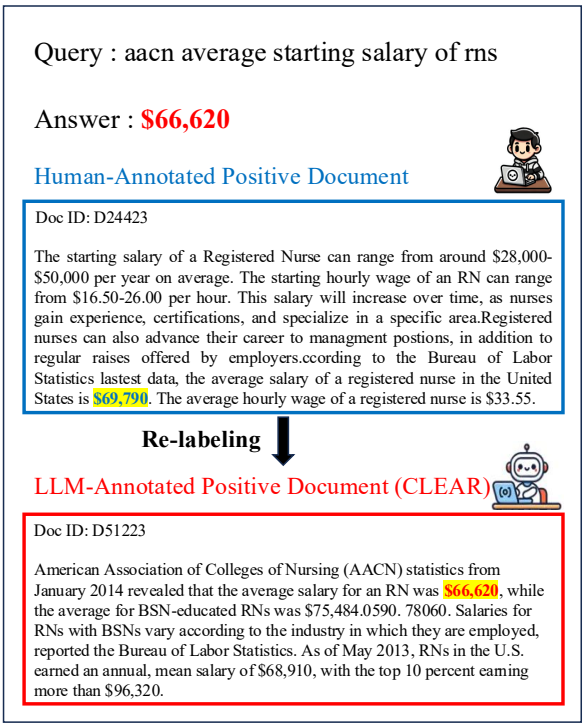

Figure 1: An example from the MS MARCO dataset comparing human-annotated and LLM-annotated positive documents for the query "AACN average starting salary of RNs." The ground truth answer is $66,620. The human-annotated document provides general salary ranges for registered nurses but does not explicitly mention the exact answer. In contrast, the LLM-annotated document explicitly states the answer, referencing AACN statistics.

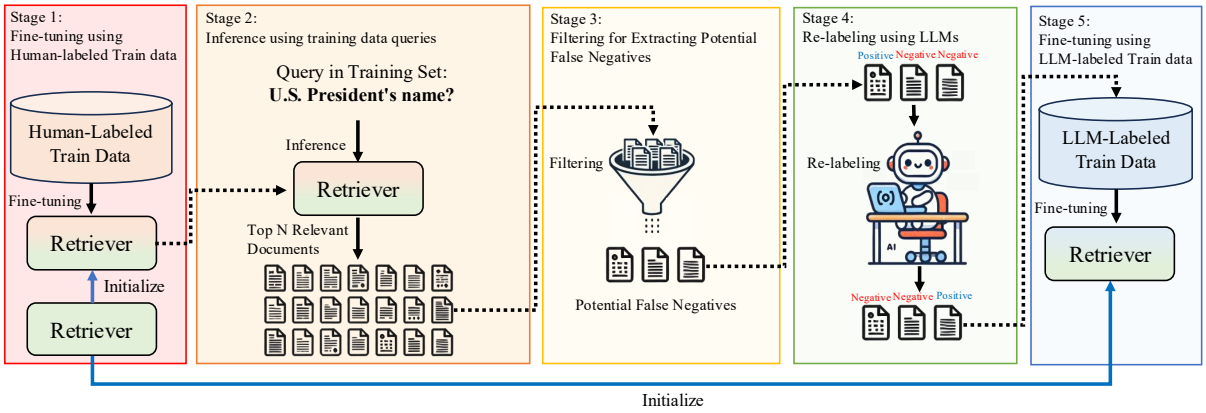

Figure 2: The CLEAR pipeline is designed to improve retriever training through LLM-based re-labeling. The process consists of five stages: (1) fine-tuning an initial retriever using human-labeled training data, (2) running inference on training queries, (3) filtering potential false negatives, (4) LLM-based re-labeling of retrieved documents, and (5) fine-tuning the retriever with the LLM-labeled dataset. The CLEAR framework enhances retrieval quality by correcting label errors and refining training data.

retrieval (IR) systems (Klie et al., 2023; Agro and Aldarmaki, 2023).

Figure 1 illustrates a comparison between a mislabeling example by a human annotator in the MS MARCO dataset (Bajaj et al., 2016) and the corrected labeling generated by the proposed CLEAR method. Whereas the human-annotated passage does not explicitly contain the correct answer, the passage labeled by CLEAR clearly provides a precise and direct response to the query.

Incorrect labels can significantly distort the evaluation of retrieval models and impede the training of optimal models. Therefore, ensuring label accuracy is a critical prerequisite for the development of reliable and effective retrieval models. To address this issue, we take inspiration from the human process of labeling documents. In manual annotation, annotators commonly select as positive the document that most clearly provides the correct answer to a given query among those retrieved by a search model. The labeling process can be interpreted as an assessment of how explicitly each document presents the answer to the query. Building on this insight, we propose CLEAR, a novel pipeline that leverages LLMs to efficiently and accurately identify positive documents. CLEAR is designed to replicate the human labeling process while remaining model-agnostic and broadly applicable across diverse retrieval and LLM configurations.

Recent advances in information retrieval have increasingly emphasized the use of hard negative documents to enhance model performance (Zhan et al., 2021; Xiong et al., 2020; Karpukhin et al., 2020; Ren et al., 2021). However, in real-world scenarios, a query is typically associated with multiple relevant documents rather than a single positive instance. This observation underscores the importance of identifying and leveraging a diverse set of positive documents during training (Dong et al., 2024; Xu et al., 2019). In this study, we investigate several training strategies designed to effectively incorporate multiple positive documents and conduct systematic experiments to evaluate their impact on retrieval performance. Our findings highlight the critical roles of both the quality and diversity of positive samples, offering practical insights into the development of more robust learning paradigms for information retrieval models.

Our contributions are summarized as follows:

1. We introduce CLEAR, a novel pipeline that leverages LLMs to automatically correct noisy labels in existing information retrieval datasets and construct diverse sets of high quality positive documents. CLEAR emulates the human annotation process to enhance both the accuracy and reliability of training data, and it is designed to be readily applicable across different models and retrieval settings.

2. While prior research has predominantly focused on enhancing retrieval performance through the selection of hard negative documents, we underscore the complementary role of positive document quality and diversity. We propose several training strategies for the effective utilization of multiple positive documents and demonstrate their efficacy through systematic empirical evaluation.

3. We evaluate the effectiveness of CLEAR

across a range of widely used benchmark datasets, including MS MARCO Passage, MS MARCO Document, Natural Questions, and TriviaQA. Experimental results show that CLEAR consistently achieves competitive performance relative to strong baselines across all datasets.

2 Our Method

Figure 2 presents the overall pipeline of the proposed CLEAR methodology. The CLEAR framework consists of five sequential stages, each of which is described in detail in this section. We particularly emphasize the process of re-labeling Information Retrieval (IR) datasets utilizing LLMs, along with the training strategies designed to effectively leverage the re-labeled data for improving retrieval model performance.

2.1 Stage 1: Fine-tuning using Human-labeled Train data

In the first stage, we fine-tune a dense retrieval (DR) model using human-labeled data. Specifically, the DR model is optimized via in-batch negative sampling and the InfoNCE loss (Oord et al., 2018; Bertram et al., 2024; Wu et al., 2021). Contrastive learning (CL), a widely adopted framework for training DR models, encourages the model to effectively distinguish positive document pairs from negative ones. The model is trained to minimize the following InfoNCE loss:

$$\mathcal{L}_{\text{CL}} = -\log\left(\frac{\exp(\text{sim}(q, d^+))}{\exp(\text{sim}(q, d^+)) + \sum_{j=1}^{N}\exp(\text{sim}(q, d_j^-))}\right) \tag{1}$$

where q denotes the input query, d^+ represents a positive document relevant to the query, d^- indicates a negative document, and $\text{sim}(\cdot,\cdot)$ denotes the dot product between the embeddings of the query and the document.

This initial step establishes the foundation for the subsequent LLM-based automatic re-labeling process, thereby improving both the effectiveness and stability of the CLEAR framework.

2.2 Stage 2: Inference using training data queries

In the second stage, we perform inference over the entire document collection using the dense retrieval (DR) model fine-tuned in Stage 1. For each query in the training set, the model retrieves the top-N candidate documents with the highest predicted relevance scores.

Let D denote the set of documents retrieved during Stage 2 inference. We formally define D as:

$$D = \{d_{i,1}, d_{i,2}, \ldots, d_{i,N}\}_{i=1}^{m} \tag{2}$$

where m is the number of training queries, and for each query q_i, the documents $d_{i,j}$ are the top-N documents retrieved by the DR model based on their similarity scores.

2.3 Stage 3: Filtering for Extracting Potential False Negatives

Re-labeling all top-N documents retrieved in Stage 2 using LLMs is both computationally intensive and time-consuming. To mitigate this challenge, we selectively extract candidate documents that are highly likely to be true positives. We refer to these candidate documents as *Potential False Negatives* (Moreira et al., 2024).

This filtering strategy is based on prior work (Moreira et al., 2024), which demonstrated that retriever performance can be enhanced through more effective hard negative mining.

In particular, the study showed that carefully excluding potential false negatives from the negative set yields substantial performance improvements, as the inclusion of true positives among negatives can reduce training quality.

Unlike prior studies that primarily focus on eliminating potential false negatives from the negative set, our approach seeks to identify and extract documents that are likely to be positive instances.

To extract potential false negatives, we dynamically determine a similarity threshold based on the score s^+ between the query and its corresponding human-labeled positive document. Specifically, the threshold is defined as follows:

$$\text{Threshold} = \tau \cdot s^+ \tag{3}$$

Following prior work (Moreira et al., 2024), the threshold τ is empirically set to 0.95, as this value has been shown to be effective in filtering potential false negatives.

The similarity scores between each query and its retrieved documents are formally defined as:

$$S = \{s_{i,1}, s_{i,2}, \ldots, s_{i,N}\}_{i=1}^{m} \tag{4}$$

Based on the similarity scores $s_{i,j}$, each document $d_{i,j}$ is classified according to the following criteria:

$$d_{i,j} = \begin{cases} \text{Potential False Negative}, & \text{if } s_{i,j} > \text{Threshold} \\ \text{Hard Negative}, & \text{otherwise} \end{cases} \tag{5}$$

for $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, N$

We define the final set of Potential False Negative (PFN) documents as follows:

$$\text{PFN} = \left\{ d^*_{i,1}, d^*_{i,2}, \ldots, d^*_{i,k-1}, d^+_{i,k} \right\}_{i=1}^{m} \tag{6}$$

where $\left\{ d^*_{i,1}, \ldots, d^*_{i,k-1} \right\}$ represents the documents identified as Potential False Negatives, and $d^+_{i,k}$ is the human-labeled positive document for query i.

By incorporating the Potential False Negatives alongside the human-labeled positive documents, the overall reliability of the training set is enhanced. The filtered PFN documents are subsequently forwarded to the next stage, where they are re-labeled using a large language model (LLM). This selective filtering strategy substantially reduces computational overhead compared to re-labeling all retrieved candidates.

2.4 Stage 4: Re-labeling using LLMs

In the fourth stage, we re-label the Potential False Negative documents identified in Stage 3 by leveraging LLMs. We utilize the LLM to generate an answer based on each Potential False Negative document and subsequently compute a confidence score that measures how accurately the LLM generates the correct answer.

Specifically, for each query q_i, we construct an input set comprising pairs of PFN documents from Stage 3 and the corresponding answer a_i. The input set is formally defined as follows:

$$\mathcal{I} = \left\{ (q_i, d^*_{i,1}, a_i), \ldots, (q_i, d^+_{i,k}, a_i) \right\}_{i=1}^{m} \tag{7}$$

Each input tuple is provided to the LLM, which computes a document-specific confidence score cs based on the model's predicted output distribution:

$$cs = 1 - d\left(\text{GT},\ p(y \mid T, q, d)\right) \tag{8}$$

where, T denotes the prompt template, and GT represents a binary vector that indicates the ground-truth answer tokens. The term $p(y \mid T, q, d)$ refers to the LLM's predicted probability distribution over the output sequence y, conditioned on the prompt T, query q, and document d.

The function $d(\cdot, \cdot)$ computes the distance between the distributions using the length-normalized L_2 norm, defined as follows:

$$d(p, q) = \sqrt{\frac{1}{L} \sum_{h=1}^{L} (p_h - q_h)^2} \tag{9}$$

where p_h and q_h represent the h-th elements of the probability distributions p and q, respectively, and L is the number of tokens in the ground-truth answer. This normalization ensures that the distance measure remains consistent across different sequence lengths.

A higher confidence score indicates that the document allows the LLM to predict the answer with greater accuracy. The complete set of confidence scores is formally defined as:

$$\mathcal{C} = \{ cs_{i,1}, \ldots, cs_{i,k} \}_{i=1}^{m} \tag{10}$$

2.5 Stage 5: Fine-tuning Using LLM-labeled Train Data

In the fifth stage, we propose several re-labeling strategies utilizing the confidence scores $\mathcal{C}$ obtained in Stage 4. Furthermore, we detail the corresponding fine-tuning methodologies designed to effectively exploit the re-labeled samples for improved model performance.

2.5.1 Fine-tuning Using Only LLM-labeled Data

The first strategy focuses on fine-tuning the model exclusively using positive documents that have been re-labeled by the LLM.

For each query, we select the document with the highest confidence score from the candidate set $\mathcal{C}$ and designate it as the new positive document. Formally, this selection is defined as follows:

$$d_i^{(\text{LLM+})} = \arg\max_k \ cs_{i,k}, \quad \forall i \in \{1, \ldots, m\} \tag{11}$$

where $d_i^{(\text{LLM+})}$ denotes the newly selected positive document, determined according to the confidence scores assigned by the LLM. Subsequently, the model is fine-tuned on these re-labeled documents using the InfoNCE loss function as defined in Equation (1).

2.5.2 Augmenting Human-Labeled Data with LLM-Labeled Data

The second strategy entails augmenting human-labeled data with data annotated by an LLM to enhance model performance. The primary motivation for this approach is to address potential omissions or inaccuracies in the human annotations, thereby improving both the quality and the diversity of the dataset.

To this end, we construct an augmented dataset, denoted as D_{Aug} by combining the human-labeled dataset D_{Human} with the LLM-labeled dataset D_{LLM}. The human-labeled dataset is formally defined as:

$$D_{\text{Human}} = \left\{ (q_i, d_i^{(\text{Human+})}, d_i^-) \right\}_{i=1}^{m} \quad (12)$$

The LLM-labeled dataset is defined as:

$$D_{\text{LLM}} = \left\{ (q_i, d_i^{(\text{LLM+})}, d_i^-) \right\}_{i=1}^{z}, \quad z \leq m \quad (13)$$

where $d_i^{(\text{Human+})}$ and $d_i^{(\text{LLM+})}$ represent the positive documents selected by the human annotators and the LLM, respectively. To avoid redundancy, any sample in D_{LLM} that overlaps with the human-labeled positives in D_{Human} is excluded.

The final augmented dataset D_{Aug} is defined as:

$$D_{\text{Aug}} = D_{\text{Human}} \cup D_{\text{LLM}} \quad (14)$$

Subsequently, the model is fine-tuned on the augmented dataset using the InfoNCE loss function Equation (1).

2.5.3 Joint Training of Human-Labeled and LLM-Labeled Data via Confidence Thresholding

The third strategy is based on the hypothesis that a single query may correspond to multiple positive documents. Under this assumption, all documents whose confidence scores exceed a predefined threshold ϕ are regarded as positive examples. Formally, the positive document assignment is defined as follows:

$$d_{i,j} = \begin{cases} \text{Labeled as Positive}, & \text{if } cs_{i,j} > \phi \\ \text{Labeled as Negative}, & \text{otherwise} \end{cases} \quad (15)$$

where ϕ denotes the predefined confidence threshold, and $cs_{i,j}$ is the confidence score of the j-th document for query i. The dataset D_{LLM}, comprising up to u positive documents selected based on the confidence threshold, is formally defined as:

$$D_{\text{LLM}} = \left\{ (q_i, d_{i,1}^{(\text{LLM+})}, \ldots, d_{i,u}^{(\text{LLM+})}, d_i^-) \right\}_{i=1}^{m} \quad (16)$$

- **Averaging multi-positive (AMP) loss**

We introduce a novel loss function, termed *Averaging Multi-Positive (AMP) Loss*, which is specifically designed to facilitate effective learning from multiple positive documents. AMP Loss promotes balanced optimization by assigning equal importance to all positive samples. Assuming a batch size of 1 for simplicity, the AMP Loss is formally defined as follows:

$$\mathcal{L}_{\text{AMP}} = -\frac{1}{u} \sum_{i=1}^{u} \log \left(\frac{\exp(\text{sim}(q, d_i^+))}{\exp(\text{sim}(q, d_i^+)) + \sum_{j=1}^{N} \exp(\text{sim}(q, d_j^-))} \right) \quad (17)$$

where u is the number of positive documents exceeding the threshold ϕ, d_i^+ represents the i-th positive document, and d_j^- denotes a negative document.

- **Confidence-guided multi-positive (CMP) loss**

Although AMP Loss assigns equal weights to all positive samples, this approach may not be optimal because some documents provide much more relevant or clearer answers to the query than others.

To address this limitation, we propose the *Confidence-Guided Multi-Positive (CMP) Loss*, which assigns dynamic weights to positive samples based on their confidence scores predicted by an LLM.

The CMP loss is formally defined as follows:

$$\mathcal{L}_{\text{CMP}} = -\sum_{i=1}^{u} w_i \times \log \left(\frac{\exp(\text{sim}(q, d_i^+))}{\exp(\text{sim}(q, d_i^+)) + \sum_{j=1}^{N} \exp(\text{sim}(q, d_j^-))} \right) \quad (18)$$

where the confidence-based weight w_i is given by:

$$w_i = \frac{\exp(cs_i)}{\sum_{k=1}^{u} \exp(cs_k)} \quad (19)$$

In this formulation, each positive sample's contribution to the loss is modulated by its associated confidence score, allowing the model to more effectively leverage soft supervision signals generated by the LLM.

3 Experimental Setup

3.1 Comparison Systems

To assess the effectiveness of our proposed method, we conduct a comparative evaluation against the following three representative dense retrieval models:

- **DPR :** DPR adopts a dual-encoder architecture that independently encodes queries and documents (Karpukhin et al., 2020). The similarity between a query and a document is measured via the dot product of their respective embeddings.

- **CoCondenser:** CoCondenser builds upon the Condenser model by enhancing pretraining with unsupervised learning techniques (Gao and Callan, 2021). A central contribution is the introduction of *corpus-level contrastive*

Models	Natural Questions		TriviaQA		MS-MARCO (Pas)		MS-MARCO (Doc)	
	R@5	R@20	R@5	R@20	R@5	R@20	R@5	R@20
DPR - Human-only (Pos=1, InfoNCE)	65.6	77.5	69.4	78.1	40.4	61.6	40.1	65.4
DPR - LLM-only (Pos=1, InfoNCE)	66.1	79.3	69.5	78.4	40.5	61.7	40.0	65.4
DPR - Human+LLM Aug (Pos=1, InfoNCE)	67.1	80.2	70.6	79.9	41.4	62.9	41.1	66.2
DPR - Human+LLM Thresh (Pos=N, AMP)	67.6	80.5	71.1	81.1	41.9	63.7	42.0	67.0
DPR - Human+LLM Thresh (Pos=N, CMP) **[CLEAR]**	68.8 **(+3.2%)**	81.1 **(+3.6%)**	72.8 **(+3.4%)**	81.6 **(+3.5%)**	42.4 **(+2.0%)**	64.2 **(+2.1%)**	42.5 **(+2.4%)**	67.5 **(+2.5%)**
CoCondenser - Human-only (Pos=1, InfoNCE)	72.8	80.1	73.4	80.2	45.0	68.9	43.4	71.1
CoCondenser - LLM-only (Pos=1, InfoNCE)	73.0	80.9	73.4	80.6	45.2	68.2	43.6	71.3
CoCondenser - Human+LLM Aug (Pos=1, InfoNCE)	74.1	81.2	74.8	81.1	45.9	68.6	44.0	72.0
CoCondenser - Human+LLM Thresh (Pos=N, AMP)	74.7	82.6	75.5	82.6	46.9	69.1	44.9	72.5
CoCondenser - Human+LLM Thresh (Pos=N, CMP) **[CLEAR]**	75.7 **(+2.9%)**	82.9 **(+2.8%)**	76.6 **(+3.2%)**	83.3 **(+3.1%)**	47.1 **(+2.1%)**	70.1 **(+1.9%)**	45.5 **(+2.1%)**	73.5 **(+2.4%)**
DRAGON - Human-only (Pos=1, InfoNCE)	71.5	81.8	73.9	82.3	53.1	74.7	48.1	74.3
DRAGON - LLM-only (Pos=1, InfoNCE)	71.9	82.1	74.1	82.4	53.7	74.9	48.6	74.9
DRAGON - Human+LLM Aug (Pos=1, InfoNCE)	72.5	82.7	75.4	84.0	54.0	75.5	49.2	75.4
DRAGON - Human+LLM Thresh (Pos=N, AMP)	72.9	83.6	75.7	84.2	54.1	76.0	49.5	75.9
DRAGON - Human+LLM Thresh (Pos=N, CMP) **[CLEAR]**	**73.9** **(+2.4%)**	**84.4** **(+2.6%)**	**76.1** **(+2.2%)**	**84.6** **(+2.3%)**	**54.9** **(+1.8%)**	**76.6** **(+1.9%)**	**50.1** **(+2.0%)**	**76.6** **(+2.3%)**

Table 1: Performance comparison of various retrieval models across four datasets, evaluated using Recall@5 and Recall@20 metrics. Models are trained with InfoNCE Loss (InfoNCE), Averaging Multi-Positive Loss (AMP), and Confidence-guided Multi-Positive Loss (CMP). Our proposed method, CLEAR, which leverages LLM-generated positives selected based on confidence scores, consistently outperforms the baselines. Percentage improvements over the baselines are reported in parentheses.

learning, which strengthens the semantic representations of documents and significantly improves retrieval performance across various benchmarks.

- **DRAGON:** DRAGON advances dense retrieval by employing aggressive data augmentation strategies, including both *query augmentation* and *label augmentation*, to generate a broader diversity of training examples (Lin et al., 2023).

3.2 LLMs Used for Re-labeling

To generate confidence scores and re-label training samples, we leverage a diverse set of LLMs with varying scales and architectural characteristics. Specifically, we utilize LLaMA-3.1-70B, LLaMA-3.1-8B (Touvron et al., 2023), EXAONE 3.5-32B (Research et al., 2024), Gemma-7B (Team et al., 2024), and Qwen 2.5-7B (Yang et al., 2024). Among these models, we conduct our experiments using LLaMA-3.1-70B.

3.3 Training

For fair comparison, we apply consistent training configurations and hyperparameters across all baseline models and our proposed method. All experiments are conducted on a single NVIDIA A100-SXM4-40GB GPU.

To ensure efficient training and stable evaluation, we adopt the batch size recommended for each model, following the configurations specified in their original implementations. The number of hard negatives is set to one in all experiments to minimize performance variance introduced by different negative sampling strategies.

During training, we select the checkpoint that achieves the highest validation score for final evaluation. This procedure ensures that each model is assessed at its optimal performance level under a consistent training protocol.

4 Experiments

Table 1 summarizes the retrieval performance of our proposed CLEAR pipeline compared to strong

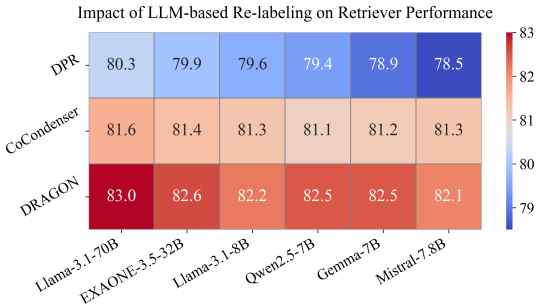

Figure 3: Retriever performance across different LLMs on the Natural Questions dataset, measured by Recall@20. The heatmap compares the retrieval effectiveness of three retrievers when paired with various LLMs, including Llama-3.1-70B, EXAONE-3.5-32B, and others. Higher recall scores are indicated in red, while lower scores are in blue.

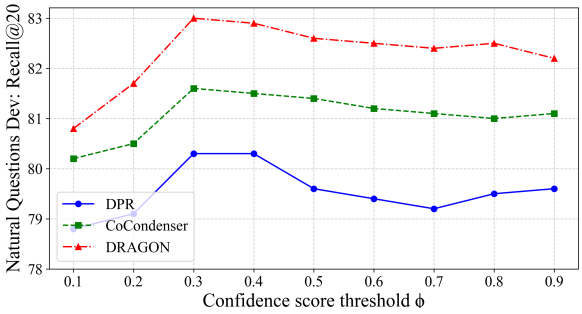

Figure 4: Impact of confidence score threshold ϕ on Recall@20 for the Natural Questions development set. The plot compares the performance of three models as the confidence threshold varies from 0.1 to 0.9.

baselines across four benchmark datasets: Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and MS MARCO (Bajaj et al., 2016). All datasets used in our study are in English and primarily cover web-based passages and open-domain questions. MS MARCO consists of real anonymized Bing queries and passages retrieved from web documents. Natural Questions consists of real, anonymized queries issued to the Google search engine, paired with Wikipedia articles retrieved at the time of the query. Our re-labeled dataset inherits these properties. For the Natural Questions and TriviaQA datasets, we use the same train/dev/test splits as provided in the original benchmark releases (Karpukhin et al., 2020). For MS MARCO, we use the publicly available dataset without any modification. Retrieval performance is measured using Recall@5 and Recall@20 metrics.

4.1 Effect of LLM-based Labeling (LLM-only)

Models trained on labels generated by LLMs consistently outperform those trained on human-annotated labels. Our analysis shows that approximately 10% of the documents were re-labeled by the LLMs, while the remaining 90% exhibited identical labels between human annotators and LLMs. These results indicate that LLMs compensate for human labeling errors or identify more appropriate positive documents. The fact that the majority (90%) of labels remain consistent suggests that LLMs largely preserve high-quality human judgments, while the re-labeled 10% likely capture edge cases such as relevant documents missed by annotators due to fatigue, subjective interpretation, or limited context.

4.2 Impact of Augmenting Data with Both Human and LLM Labels (Human+LLM)

Augmenting human-annotated data with labels generated by LLMs consistently improves retrieval performance. This finding suggests that human-labeled and LLM-labeled documents serve complementary functions, jointly contributing to enhanced retrieval effectiveness.In many cases, both human-annotated and LLM-labeled documents can be considered valid positive examples, reflecting the multiplicity of relevance judgments.These findings underscore the value of combining human and LLM supervision to construct richer and more semantically diverse training signals, ultimately leading to more robust retrieval models.

4.3 Effectiveness of Multi-Positive Training (Joint Training, AMP Loss)

Training with multiple positive documents consistently outperforms training with a single positive document. These findings indicate that leveraging multiple positive examples facilitates more stable and robust model learning. We hypothesize that this improvement stems from the increased diversity and coverage provided by multi-positive supervision. In contrast to single-positive training, where the model is optimized to match a narrow view of relevance, multi-positive training exposes the model to a wider semantic spectrum of valid answers. This helps the model generalize better to unseen queries by reducing overfitting to a limited set of lexical or structural patterns. Additionally, averaging over multiple positives during loss com-

putation smooths the learning signal and mitigates the influence of outlier examples, further contributing to optimization stability and performance robustness.

4.4 Effectiveness of Confidence-Guided Weighting (Joint Training, CMP Loss)

In multi-positive training, uniformly assigning weights to all positive documents may not always yield optimal performance. This is because not all positive documents hold the same importance with respect to the query. To address this issue, we employ a confidence-guided weighting strategy that dynamically adjusts the contribution of each positive document based on confidence scores provided by the LLM. This strategy is particularly beneficial in scenarios where some LLM-labeled positives are only weakly relevant or noisy. By down-weighting low-confidence examples, the model can avoid overfitting to uncertain supervision signals and allow positive documents with higher confidence scores to exert a greater influence during training.

4.5 Comparative Analysis of Retriever Performance with Various LLM Labelers

Figure 3 illustrates the impact of LLM-based relabeling on the training of retrieval models. In the proposed framework, a LLM receives a query and a document as input and generates a response indicating whether the document contains the correct answer. LLMs with larger parameter sizes possess greater parametric knowledge, enabling them to generate more accurate and reliable labels.

Experimental results demonstrate that retrieval performance improves with the scale of the LLM's parameters. Notably, the DRAGON model achieved the highest Recall@20 when trained with labels generated by LLaMA-3.1-70B, closely followed by EXAONE-3.5-32B.

In contrast, LLMs with relatively smaller parameter sizes—such as LLaMA-3.1-8B, Qwen2.5-7B, and Gemma-7B—exhibited comparable performance levels, whereas Mistral-7B consistently yielded the lowest Recall@20 scores across all retrieval models. This suggests that lower-quality answers generated by smaller LLMs can degrade label quality and, in turn, negatively impact downstream training performance.

This trend was consistently observed across different retrieval models, including DRAGON, CoCondenser, and DPR. These findings underscore the importance of selecting a sufficiently large LLM for re-labeling, as high-quality supervision from high-capacity models can substantially enhance retrieval effectiveness.

4.6 Impact of Confidence Score Threshold on Retrieval Performance

Figure 4 presents the impact of the confidence score threshold (ϕ) on retrieval performance. The figure compares Recall@20 across three models—DPR, CoCondenser, and DRAGON—under varying threshold values, highlighting how filtering based on LLM-generated confidence scores affects retrieval quality.

Overall, increasing the confidence score threshold leads to a decrease in Recall@20. This trend indicates that overly aggressive filtering based on high confidence scores may inadvertently exclude valuable positive samples, thereby impairing retrieval effectiveness.

The highest performance is observed at $\phi = 0.3$, suggesting that removing low-confidence, potentially noisy positive samples can contribute to improved model training. At $\phi = 0.3$, an average of 3.5 positive documents are retained per query. These results suggest that maintaining a lower confidence threshold, which allows for a greater diversity of positive documents during training, can further enhance retrieval performance.

5 Conclusion

In this work, we propose CLEAR, a novel pipeline that improves the quality of IR training datasets via LLM-based re-labeling. By correcting noisy labels and identifying diverse, high-quality positives, CLEAR enhances both the accuracy and coverage of supervision.

Experiments on four benchmark datasets show that CLEAR consistently improves retrieval performance across multiple retrievers. We also demonstrate that confidence-guided weighting in multi-positive training stabilizes optimization and enhances generalization.

These results underscore the value of LLMs as effective tools for constructing reliable IR datasets and motivate future research on automated label refinement and soft-supervision in retrieval tasks.

6 Limitations

Answer Dependency CLEAR relies on the LLM-generated answers to compute confidence scores for document re-labeling. This approach

inherently assumes the accuracy of each generated question–answer pair. However, if an answer is incorrect, the resulting confidence score may lead to erroneous re-labeling of documents, thereby propagating inaccuracies within the dataset.

Heuristic Sensitivity The use of fixed thresholds and heuristic-based filtering may lead to suboptimal performance in domains with significantly different distributions. These manually tuned parameters, while effective on our validation set, are unlikely to transfer robustly to new domains, specialized corpora, or query styles that deviate from open-domain benchmarks.

Domain and Language Constraints Experiments are conducted only on English-language benchmarks in open-domain IR. It remains unclear whether CLEAR can be effectively applied to other languages, low-resource settings, or highly domain-specific corpora such as legal or medical text.

7 Ethical Considerations

Data Source Transparency We use only publicly available datasets—MS MARCO, Natural Questions, TriviaQA, and others—which were released for academic use and contain no personally identifying information. No additional human data was collected or annotated.

Bias and Fairness Concerns While CLEAR aims to improve label quality, it inherits potential biases from both the original human annotations and the LLM used for re-labeling. For example, LLM-generated answers may reinforce patterns present in web-scale pretraining data, leading to unintentional biases in re-labeled datasets.

Responsible Use Our re-labeled data and pipeline are intended strictly for academic research. Practitioners adopting CLEAR should be cautious about unintended consequences of relying on LLM-generated pseudo-labels, especially in sensitive application domains. Future work should explore mechanisms to verify or calibrate LLM-generated outputs for better safety and transparency.

References

Maha Tufail Agro and Hanan Aldarmaki. 2023. Handling realistic label noise in bert text classification. *arXiv preprint arXiv:2305.16337*.

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

Timo Bertram, Johannes Fürnkranz, and Martin Müller. 2024. Contrastive learning of preferences with a contextual infonce loss. *arXiv preprint arXiv:2407.05898*.

Derek Chong, Jenny Hong, and Christopher D Manning. 2022. Detecting label errors by using pre-trained language models. *arXiv preprint arXiv:2205.12702*.

Hengkui Dong, Xianzhong Long, and Yun Li. 2024. Rethinking samples selection for contrastive learning: Mining of potential samples. *Knowledge-Based Systems*, 299:111979.

Luyu Gao and Jamie Callan. 2021. Unsupervised corpus aware language model pre-training for dense passage retrieval. *arXiv preprint arXiv:2108.05540*.

Biyang Guo, Songqiao Han, Xiao Han, Hailiang Huang, and Ting Lu. 2021. Label confusion learning to enhance text classification models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12929–12936.

Hui Guo, Boyu Wang, and Grace Yi. 2023. Label correction of crowdsourced noisy annotations with an instance-dependent noise transition model. *Advances in Neural Information Processing Systems*, 36:347–386.

Emily Jamison and Iryna Gurevych. 2015. Noise or additional information? leveraging crowdsource annotation item agreement for natural language tasks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 291–297.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

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

Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. 2023. Annotation error detection: Analyzing the past and present for a more coherent future. *Computational Linguistics*, 49(1):157–198.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. *arXiv preprint arXiv:2302.07452*.

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. *arXiv preprint arXiv:2407.15831*.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. 2021. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11).

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

LG Research, Soyoung An, Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Gerrard Jeongwon Jo, and 1 others. 2024. Exaone 3.5: Series of large language models for real-world use cases. *arXiv preprint arXiv:2412.04862*.

Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622.

Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Chuhan Wu, Fangzhao Wu, and Yongfeng Huang. 2021. Rethinking infonce: How many negative samples do you need? *arXiv preprint arXiv:2105.13003*.

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

Miao Xu, Bingcong Li, Gang Niu, Bo Han, and Masashi Sugiyama. 2019. Revisiting sample selection approach to positive-unlabeled learning: Turning unlabeled data into positive rather than negative. *arXiv preprint arXiv:1901.10155*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

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

Han Zhang, Yazhou Zhang, Jiajun Li, Junxiu Liu, and Lixia Ji. 2025. A survey on learning with noisy labels in natural language processing: How to train models with label noise. *Engineering Applications of Artificial Intelligence*, 146:110157.

A Appendix

Model : DPR	Llama-3.1-70B	Llama-3.1-8B
Threshold : 0.95	66.1	**65.4**
Threshold : 0.90	66.3	65.1
Threshold : 0.85	66.3	65.0
Threshold : 0.80	**66.5**	64.7
Threshold : 0.75	66.4	64.5
Threshold : 0.70	66.2	64.6

Table 2: Recall@5 for DPR LLM-only under different thresholds on the NQ dataset

A.1 Impact of Filtering Threshold on Retrieval Performance

In Stage 3, we apply filtering to identify potential positive candidates based on a thresholding strategy. Since re-labeling all retrieved documents with LLMs is computationally expensive and often unnecessary, we first narrow down the candidate pool through filtering to reduce labeling overhead. Specifically, we follow the hyperparameter settings proposed in (Moreira et al., 2024), which

suggest using a similarity threshold to extract potential false negatives based on their proximity to human-labeled positives. The original study (Moreira et al., 2024) reports that setting the threshold to 0.95 is particularly effective for removing hard negatives that are semantically close to positives. We then apply LLM-based re-labeling only to the filtered candidate documents.

To validate its applicability in our framework, we re-experimented with varying threshold values. Our results show that higher thresholds tend to slightly improve performance, especially when stronger LLMs are used for document re-labeling, as they are more capable of correctly identifying true positives from a larger pool of candidates. This suggests that LLMs with larger parameter counts exhibit better semantic understanding, enabling more accurate re-labeling decisions even when the filtering threshold is relaxed and a broader range of candidate documents is considered.