applied sciences

MDPI

*Article*

# Towards Robust Neural Rankers with Large Language Model: A Contrastive Training Approach

**Ziyang Pan**, **Kangjia Fan, Rongyu Liu and Daifeng Li ***

School of Information Management, Sun Yat-sen University, Guangzhou 510275, China
* Correspondence: lidaifeng@mail.sysu.edu.cn

**Abstract:** Pre-trained language model-based neural rankers have been widely applied in information retrieval (IR). However, the robustness issue of current IR models has not received sufficient attention, which could significantly impact the user experience in practical applications. In this study, we focus on the defensive ability of IR models against query attacks while guaranteeing their retrieval performance. We discover that improving the robustness of IR models not only requires a focus on model architecture and training methods but is also closely related to the quality of data. Different from previous research, we use large language models (LLMs) to generate query variations with the same intent, which exhibit richer and more realistic expressions while maintaining consistent query intent. Based on LLM-generated query variations, we propose a novel contrastive training framework that substantially enhances the robustness of IR models to query perturbations. Specifically, we combine the contrastive loss in the representation space of query variations with the ranking loss in the retrieval training stage to improve the model's ability to understand the underlying semantic information of queries. Experimental results on two public datasets, WikiQA and ANTIQUE, demonstrate that the proposed contrastive training approach effectively improves the robustness of models facing query attack scenarios while outperforming baselines in retrieval performance. Compared with the best baseline approach, the improvements in average robustness performance of Reranker IR models are 24.9%, 26.5%, 27.0%, and 75.0% on WikiQA and 8.7%, 1.9%, 6.3%, and 13.6% on ANTIQUE, in terms of the MAP (Mean Average Precision), MRR (Mean Reciprocal Rank), nDCG@10 (Normalized Discounted Cumulative Gain) and P@10 (Precision), respectively.

**Keywords:** contrastive learning; neural rankers; robustness; large language model

check for updates

**Citation:** Pan, Z.; Fan, K.; Liu, R.; Li, D. Towards Robust Neural Rankers with Large Language Model: A Contrastive Training Approach. *Appl. Sci.* **2023**, *13*, 10148. https://doi.org/10.3390/app131810148

Academic Editor: Valentino Santucci

Received: 31 July 2023
Revised: 1 September 2023
Accepted: 4 September 2023
Published: 8 September 2023

## 1. Introduction

In recent years, a plethora of work has demonstrated the astonishing performance of pre-trained models (PLMs) [1–4] in natural language processing (NLP). The new research paradigm, which involves fine-tuning PLMs on self-supervised tasks for specific applications, has been rapidly adopted by researchers in the community [5]. The paradigm shift in NLP research has also brought new perspectives to information retrieval (IR) studies, which are similarly focusing on text processing. Existing research [6] has indicated that leveraging PLMs with remarkable performance on NLP tasks could significantly enhance the performance of IR models. Although these models exhibit formidable capabilities, including ranking effectiveness and efficiency, they tend to be more vulnerable to adversarial attacks compared to traditional models [7]. For example, upon the introduction of perturbations to the query, the performance of neural rankers may potentially deteriorate. These perturbations, such as word typos and synonym word replacement, are frequently encountered by IR models deployed in the wild. Despite the existence of various proposals aimed at enhancing the robustness of retrieval models, the majority of these approaches [8–11] are either task-specific or solely reliant on data augmentation. While methods for resisting perturbations at the word or sentence level (such as those addressing typos and synonym substitutions) have been well-developed, there was a notable scarcity of work focusing on query reformulations with the same intent. A key reason for this is the lack of adequate high-quality augmented data. In light of the significant performance improvement of large language models (LLMs) in generative tasks, we endeavored to employ LLMs to generate query reformulations while maintaining the same intent.

Research in computer vision (CV) has demonstrated that representations obtained under a contrastive self-supervised setting exhibited greater robustness when faced with out-of-distribution (OOD) data and image corruption [12]. In enhancing the robustness of information retrieval models with respect to query reformulations maintaining the same intent, it is fundamentally essential to focus on the robustness of query representations under perturbations from diverse distributions. Based on the definition of adversarial text [13], we posit that the semantics of various reformulations of a single query under the same intent should tend to be consistent, while the semantics between different queries should tend to diverge. Guided by this intuition, we employ a contrastive training approach, utilizing LLM-enhanced data to optimize the text representations in PLMs within information retrieval systems.

In this study, we primarily focus on the defensive ability against query attacks for IR models based on PLMs, which constitutes an essential subtask in assessing the robustness of IR. We concentrate on ensuring that information retrieval models produce stable and accurate results when faced with different query expressions under the same intent. We propose a novel contrastive learning approach to fine-tune PLM-based IR models and investigate its benefits for enhancing model robustness. Experiments are conducted on both different IR model architectures, employing multi-objective optimization loss functions. In terms of ranking, we train the neural rankers by optimizing ranking loss, which compares scores between different documents under the same query and maximizes the score for the positive candidate. Additionally, we utilize contrastive loss to optimize the representation of queries with the same intent in the latent space (i.e., within an input batch, different query expressions with the same intent should be close in the latent space, while those with different intents should be distant). The purpose of employing this contrastive loss is to stimulate the model to learn the underlying common semantic information between varying expressions of queries with the same intent, thereby enhancing the model's robustness to semantic perturbations and its ability to represent the latent intent within queries.

The main contributions of this study are summarized as follows:

- After reviewing existing text data augmentation methods [14,15], we find that current text attack methods suffer from high rule dependency and limited variation forms. In this study, we leverage the emergent general capabilities of LLMs to generate a set of high-quality query variations.
- We assume that the representations of the original query and the corresponding query variations in the latent space should be similar while being distant from other queries. To train a robust IR model against adversarial query attacks, we propose a novel contrastive training approach that contains tasks of ranking and intent alignment.
- This study conducts experiments on two public datasets, and the results demonstrate the effectiveness of the proposed training approach. Our method not only improves the query robustness of the models but also ensures that retrieval performance does not suffer significant losses, and even outperforms the original models in some metrics.

## 2. Related Work

### *2.1. Robustness of IR*

Our work primarily focuses on the ranking stage of IR, as shown in Figure 1, which entails sorting the candidate document set retrieved in response to user queries in order to meet their search requirements. From the perspective of structure design, PLM-based ranking models can be categorized into Cross Encoder (Reranker) [16] and Dense Retriever [17]. In terms of training paradigms, retrieval ranking models can be divided into pointwise (PRank [18], Ranking with Large Margin Principle [19]), pairwise (RankNet [20], LambdaRank [21], Ranking SVM [22]), and listwise (ListNet [23], AdaRank [24], LambdaMART [25]) approaches.

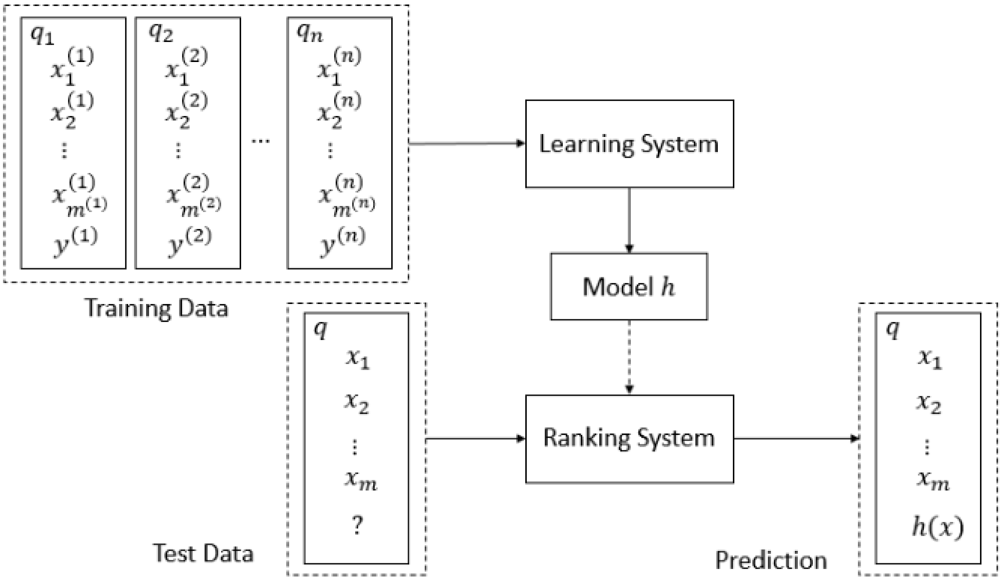

**Figure 1.** The training paradigm for IR ranking models.

As neural networks have been employed in retrieval ranking, significantly improving the performance of IR, researchers have shifted their focus towards the robustness of IR models rather than solely concentrating on their average performance metrics. Robustness is, in fact, a multi-dimensional concept, and thus Wu et al. [7] designed five robustness evaluation tasks to systematically assess the robustness of the most effective deep learning retrieval models and traditional retrieval models currently available. Although the overall experimental conclusions did not confirm that deep learning models exhibit stronger or weaker robustness compared to traditional models, the performance of neural network retrieval models in terms of robustness was not as dazzling as their performance in enhancing IR effectiveness, particularly in the presence of adversarial examples.

In addition to the robustness evaluation work [7], researchers have also dedicated work to devising algorithms to bolster the robustness of retrieval models. Arslan et al. [26] have proposed a word-weighted selection method based on chi-square value. This approach automatically selected the optimal word-weighting strategy from the candidate strategies, and their experiment showed that this method was better than any single word-weighting method in the system. Reddy et al. [27] have pointed out that the current evaluation of the retrieval model was limited to the average performance, without considering its performance in cross-domain retrieval and zero-shot learning. Therefore, they proposed a Seq2Seq model which was used to synthesize new training samples, and the experimental results on five datasets showed that the performance was better when using the synthesized data. In order to solve the problem of the annotation of large-scale training data, Prakash et al. [28] proposed a strategy to select negative samples from unlabeled data, which was to select difficult negative samples according to the score of BM25, so as to expand the original small dataset to achieve more training data and improve robustness. Zhuang et al. [10] found that the retrieval performance of existing deep retrieval models deteriorated significantly when there were spelling errors in the query. To solve this problem, they proposed a training method that can simulate multiple spelling error patterns in the query. Experiment results showed that this training method could improve the robustness of the model when there were spelling errors in the query text.

### *2.2. Contrastive Learning for IR*

Contrastive learning is a popular representation learning approach that initially demonstrated exceptional performance in computer vision [12,29]. Contrastive learning enhances the model by introducing additional contrastive tasks leveraging the contrastive pairs. The construction of positive and negative samples in NLP could often be conducted without supervision, making it widely applicable. In NLP, although preparing contrastive samples for text is not as straightforward as for images, there still exist lots of studies about contrastive learning. For instance, IS-BERT [30] has been proposed to add a CNN layer to BERT and it was trained by maximizing the mutual information (MI) between global sentence embeddings and their corresponding local context embeddings. Cheng et al. [31] constructed biased word pairs, replaced input bias words to generate augmented samples for contrastive learning, and minimized the MI between sensitive words while maximizing the MI between sample pairs to eliminate social biases in text representations. Qu et al. [32] have studied contrastive learning by employing five methods for constructing augmented samples in NLP, including back-translation, language model prediction, Mixup, cutoff (randomly replacing features with zeros), and adversarial learning, achieving improved results in various natural language understanding tasks.

In IR, some research also involves constructing positive and negative pairs for contrastive learning. For example, S-BERT [33] adopted a Siamese neural network, employing BERT [2] to obtain individual sentence representations, and treating related sentences as positive samples and other sentences within the same batch as negative samples. This approach learned text-matching tasks through contrastive learning of query-document pairs. Izacard et al. [34] have proposed a document set contrastive loss in response to issues such as the susceptibility of current retrieval models to noise, constructing contrastive samples through a local sampling of different documents, which enhances the model's cross-document set retrieval performance. Deng et al. [35] have addressed the issue of having very few labeled positive pairs in current datasets by devising a strategy that leverages the pre-trained BART model [36] to generate pseudo-relevant queries and pseudo-relevant documents, increasing the number of pseudo-positive pairs, and then constructing a new contrastive loss based on these pseudo-positive pairs.

## 3. Methodology

### *3.1. Problem Formulation*

As in previous research, we formally define the ranking problem first. Given a query $q$ and a large set of documents $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$, the task is to produce a ranked list of documents, which is as close as possible to the gold ranking of documents according to their relevance levels. For the binary label setting, the relevance level $r \in \{0, 1\}$, where 1 indicates the positive and 0 indicates the negative. There may also be multiple levels of relevance label setting, in which case the gold ranking of documents should adhere to the ranking established by the multi-level relevance labels. Once the model generates a ranked list of documents, various evaluation metrics will be employed to assess the model's retrieval performance, such as MRR, MAP, NDCG, etc. In the robustness evaluation setting, following prior research, the robust evaluation presented in this paper is also grounded in metrics related to retrieval performance, with detailed information available in the experimental section.

### *3.2. Main Framework*

We propose a novel robustness training framework, as illustrated in Figure 2, which employs high-quality query variations generated by LLMs to construct a multi-objective loss function based on contrastive learning. This framework encompasses three key modules: (1) LLM-based query generation; (2) ranking training and (3) query intent alignment. In Algorithm 1, we present the pseudocode for the overall framework, detailing the specific flow between modules.

**Algorithm 1** Pseudocode for the proposed approach

```
Input: PLM-based model, LLM (e.g., ChatGPT), Dataset $D$ (e.g., WikiQA)
Output: Fine-tuned IR model for retrieval
for each query $q_i$ in $D$ do
    Generate query variations $\{q_i^k\}$ for $q_i$ by LLM
end for
Initialize BERT with pre-trained weights
for each batch in $D$ do
    Extract input sequences $X$ and target labels $Y$
    Compute embeddings $E = \mathrm{BERT}(X)$
    Compute loss $L$ using $E$ and $Y$
    Back-propagate error to update BERT's weights
end for
return Fine-tuned IR model
```

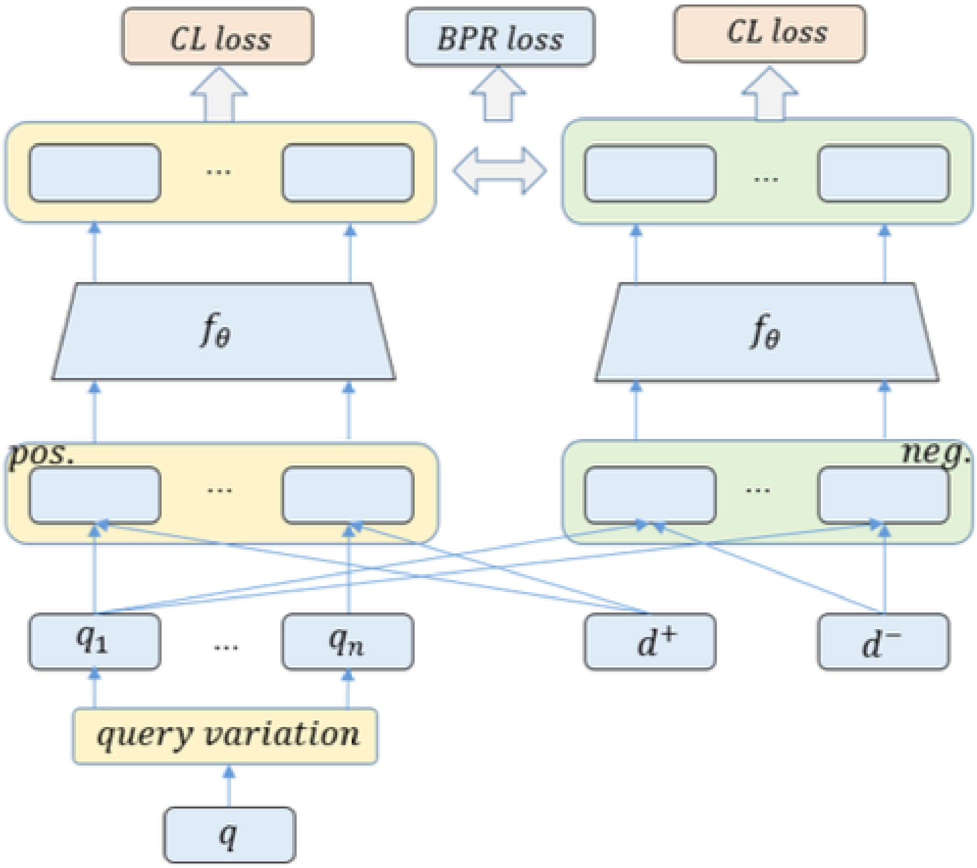

**Figure 2.** Contrastive training framework for a more robust PLM-based IR model. The original queries are used to generate different expressions with the same intent during the query variation phase, utilizing a LLM. The retrieval structure is illustrated on Reranker as an example, where BPR Loss serves as the loss function for the ranking task and Normalized Temperature-Scaled Cross Entropy Loss serves for the intent alignment task.

### 3.2.1. LLM-Based Query Generation

Recently, the remarkable performance in natural language generation (NLG) of ChatGPT [37], which was released by OpenAI, has attracted the attention of the NLP community and others worldwide. The research by Wang et al. [38] indicated that the LLM is more robust than language models usually used in IR. In existing research [14,15], adversarial text samples were mostly generated by random rules [39] or seq2seq models such as

T5 [4] and M2M100 [40]. In our experiments, we use WikiQA [41] and ANTIQUE [42], which are derived from real logs of Bing and Yahoo Answers, respectively. We provide a detailed introduction to these datasets in Section 4. As shown in Table 1, we found that many adversarial samples are inherently present in these two real-world retrieval datasets. Training with these samples may, to some extent, affect the model's ability to understand semantics and thus impact the overall retrieval performance. Surprisingly, we found that the LLM-generated query variations not only precisely captured the query intents but also exhibited a greater diversity in expression compared to the results generated by the previous method. Furthermore, LLMs possessed the capability to automatically rectify potential adversarial samples within the dataset (see Table 1). In this study, we employed the GPT-3.5-turbo API provided by OpenAI to generate query variations. The general process of data augmentation could be formally expressed as follows:

Given a query data augmentation operator $\mathcal{G}$ which consists $k$ different query variations, namely $\mathcal{G} = \{g_1, g_2, \ldots, g_k\}$. For query set $\mathcal{Q} = \{q_1, q_2, \ldots, q_M\}$, we sample each operation $g_k$ to transform a query $q_i$ to obtain different query variations for every query intention:

$$q_i^k = g_k(q_i), k \in [1, K], \tag{1}$$

where $g_k$ represents the $k$th query variation generated by the operator $\mathcal{G}$. Upon the query generated by the LLM and previous methods implemented based on Gustavo et al. [39], we conducted a comparative evaluation between them. For the task of generating query variations, it is desirable to maintain the original query's intent (i.e., high semantic similarity) while also introducing greater diversity. Therefore, we utilized six models from the Massive Text Embedding Benchmark (MTEB) [43] to obtain sentence embeddings for $q_i^k$ and calculated their semantic similarity to evaluate the semantic resemblance between the query variation and the original query. We measured the diversity of the generated query variations using Jaccard similarity, Levenshtein distance, and length of them. Detailed evaluation results could be found in Tables 2 and 3.

**Table 1.** The generated variations by ChatGPT based on the original queries in the dataset are presented in this table, which illustrates three different levels of query perturbations. The original queries from the datasets are bolded.

| Original Query | Query Variation Generated by ChatGPT | Level |
|---|---|---|
| **what are mormens?**<br>ANTIQUE: #2180086 | 1. Could you tell me about the Mormon religion?<br>2. What is the Mormon faith all about?<br>3. I'm a bit confused about Mormons - can you explain?<br>4. Can you enlighten me about the beliefs of Mormon?<br>5. I've heard about Mormon, but I'm not entirely sure what they are - can you help me understand? | **Character** |
| **what is masturbat$* * *$?**<br>ANTIQUE: #1971899 | 1. What does the term masturbation refer to?<br>2. Can you define the act of masturbation?<br>3. Could you provide an explanation of masturbation?<br>4. What is the definition of masturbation?<br>5. How would you describe masturbation? | **Word** |
| **what is Carbon 14 dating is a type of ?**<br>WikiQA: Q2325 | 1. What type of dating is Carbon-14 dating?<br>2. What is the classification of Carbon-14 dating?<br>3. Categorically speaking, what is Carbon-14 dating?<br>4. In what category does Carbon-14 dating belong?<br>5. To what genre does Carbon-14 dating pertain? | **Sentence** |

In the Table, words in red font illustrate the perturbations of specific terms in the original query and their counterparts in the query variations generated by ChatGPT. In the "original query" column ('***' represents intrinsic content within the dataset), the original queries from the dataset are shown at the top and highlighted in bold. Below, the dataset and the index of the original query within the dataset are indicated, which are annotated in blue.

**Table 2.** Semantic similarity between original query and query variations. Cosine similarity is adopted to compute the semantic similarity between queries. The higher the values, the more similar the two sentences are. The previous methods are categorized based on three different levels of perturbations: character, word, and sentence.

| | **Character-Level** | | **Word-Level** | | **Sentence-Level** | | **LLM** | |
|---|---|---|---|---|---|---|---|---|
| | **WikiQA** | **ANTIQUE** | **WikiQA** | **ANTIQUE** | **WikiQA** | **ANTIQUE** | **WikiQA** | **ANTIQUE** |
| bge-large-en [44] | 96.04% | 94.48% | 93.22% | 91.44% | 95.59% | 96.73% | 92.02% | 95.08% |
| bge-base-en [45] | 93.75% | 94.24% | 94.36% | 92.28% | 97.50% | 96.88% | 95.29% | 94.64% |
| instructor_xl [46] | 91.94% | 93.51% | 90.57% | 87.49% | 95.84% | 95.03% | 92.26% | 92.25% |
| instructor_large [46] | 96.19% | 96.82% | 95.75% | 94.17% | 98.04% | 97.58% | 96.81% | 96.93% |
| gtr-t5-xl [47] | 89.09% | 90.78% | 86.60% | 81.45% | 93.82% | 92.21% | 89.20% | 88.57% |
| gtr-t5-xxl [47] | 89.89% | 91.14% | 87.14% | 81.96% | 93.97% | 92.36% | 88.89% | 88.43% |

**Table 3.** Evaluation of the diversity of query variations in comparison to the original query. In terms of Jaccard Similarity, smaller values suggest a greater degree of deviation from the original query. Conversely, for Levenshtein Distance, larger values indicate greater diversity. The previous methods are categorized based on three different levels of perturbations: character, word, and sentence. The best result is **bold-faced**.

| | **Sentence Length** | | **Jaccard Similarity** | | **Levenshtein Distance** | |
|---|---|---|---|---|---|---|
| | **WikiQA** | **ANTIQUE** | **WikiQA** | **ANTIQUE** | **WikiQA** | **ANTIQUE** |
| Original query | 36.89 | 47.59 | - | - | - | - |
| Character-level query variation | 36.44 | 44.76 | 71.80% | 72.68% | 1.85 | 4.42 |
| Word-level query variation | 21.00 | 23.45 | 43.45% | 34.05% | 18.31 | 26.21 |
| Sentence-level query variation | 36.27 | 43.12 | 56.71% | 63.54% | 9.07 | 11.97 |
| LLM query variation | 58.48 | 68.66 | **17.72%** | **21.08%** | **34.56** | **40.07** |

It could be observed from Tables 2 and 3 that LLM-generated query variations, while maintaining semantic similarity to the original query, significantly enhanced linguistic expression diversity compared to query variations generated by previous methods. In addition, the sentence length of query variations was also included in the scope of the evaluation. We found that LLMs tended to output longer query variations. By incorporating newly generated query samples, we have expanded the original dataset $< q_i, d^+, d^- >$ into a new dataset $< q_i^k, d^+, d^- >, k \in [1, K]$. Note that, in order to ensure fairness when comparing to baselines, the original queries were still retained in the new dataset.

### 3.2.2. Ranking Training

In this section, we briefly introduce two primary structures of PLM-based ranking models: Dense Retriever and Cross Encoder (Reranker), as well as exploring how to design loss functions for the ranking task. Cross Encoder has demonstrated superior retrieval performance compared to the Dense Retriever, although it falls short in terms of computational efficiency. Therefore, we conducted experiments on both architectures to demonstrate the versatility of our approach.

**Dense Retrievers** [17,48,49] would employ a uniform BERT encoder to obtain the representations of both queries and documents:

$$H(q_i) = Encoder([CLS] \circ q_i \circ [SEP]), \tag{2}$$

$$H(d_i) = Encoder([CLS] \circ d_i \circ [SEP]). \tag{3}$$

In Equations (2) and (3), the data preprocessing adheres to the input rules of BERT, where $[CLS]$ and $[SEP]$ are special tokens, and $\circ$ represents the concatenation operation. Generally, we utilize the same encoder (in this study, we use BERT) to obtain the representations of the query and its candidate documents. This dense vector representation will be used to calculate the similarity between the query and the documents in the candidate set, denoted by $Score(q, d)$:

$$Score_{<q_i,d_i>} = sim(H(q_i), H(d_i)), \tag{4}$$

$Score(q, d)$ measures the similarity between the query and all documents within its candidate set. In Dense Retriever, the similarity function typically employs the dot product [48,49].

**Cross Encoder (Reranker)** processes both the query and document jointly in order to capture the interactive information within the text. The input sequence is constructed to obtain the hidden representation of the last layer and the representation vector corresponding to [CLS], which is:

$$h_{[CLS]}, H_{last} = Encoder([CLS] \circ q_{ids} \circ [SEP] \circ d_{ids} \circ [SEP]), \tag{5}$$

where $h_{[CLS]} \in R^{Hidden}$ is the output at the model's [CLS] position and $H_{last}$ is the output of the last hidden layer. Generally, $h_{[CLS]}$ would be leveraged as the relevance score after mapping into logits:

$$Score_{<q_i,d_i>} = MLP\big(h_{[CLS]}\big). \tag{6}$$

Similarly to Dense Retriever, $Score(q, d)$ measures the similarity between the query and all documents within its candidate set.

Upon applying the aforementioned procedure, the relevance scores $Score_{<q,d^+>}$ and $Score_{<q,d^->}$ would be obtained for each given query with respect to its corresponding positive and negative documents. Utilizing the scores, we use BPR Loss [50] to calculate the ranking loss $l$ for the ranking task in each query:

$$l(q, d^+, D^-) = -log\sigma(Score_{<q,d^+>} - Score_{<q,d^->}) \tag{7}$$

where $D^-$ is the negative documents set for the query $q$ and $\sigma$ is the sigmoid function. BPR Loss was introduced as an optimization criterion for recommendation tasks, and its underlying principles have been applied for precise representations in NLP [51,52]. In recent years, BPR has also been extensively utilized in the field of IR, encompassing areas such as question-answering systems [53], privacy protection [54], and interactive information retrieval [55]. For query variations with the same intent, we assume that the ranking results of the candidate documents should be consistent with those of the original query. Therefore, we incorporate all query variations set $Q$ belonging to the same intent into the calculations. Ultimately, we obtain the contrastive training loss, where $D^+$ is the positive set of candidate documents for the given query $q$:

$$\mathcal{L}_{BPR} = \sum_{q \in Q} \sum_{d^+ \in D^+} l(q, d^+, D^-). \tag{8}$$

### 3.2.3. Query Intent Alignment

We assume that text representations based on the same query intent should be highly similar. Due to the possibility of user-generated adversarial text in the wild, we need to enhance the defensive capabilities of language models when facing adversarial attacks. After obtaining high-quality query variations generated by ChatGPT, we prefer to use contrastive learning to align various expressions of the same query intent in the latent representation space.

For Dense Retriever, $H(q_i)$ could be directly used as the representation of query $q_i$. For Reranker, we add a BERT encoder layer to obtain the query representation:

$$v_q = Encoderlayer(H_{last}). \tag{9}$$

We use the underlying BERT as a shared module and design task-related modules on its top layer according to the ranking and alignment task, respectively. Ultimately, we obtain an underlying model that takes multiple task objectives into account through joint training with multi-task learning.

In the alignment task, the in-batch contrastive learning approach is adopted. For each input query intent $q_i, (i \in [1, batchsize])$ within the batch, assuming there are $K$ query variations $q_i^k$, we use the Normalized Temperature-Scaled Cross Entropy [12] as the query intent alignment loss:

$$\mathcal{L}_{cl} = -\sum_{q_i \in Q} \sum_{j=1}^{K} \log \left( \frac{e^{sim\left(v_{q_i}, v_{q_i^j}\right)/\tau}}{e^{sim\left(v_{q_i}, v_{q_i^j}\right)/\tau} + \sum_{q' \notin q_i} e^{sim\left(v_{q_i}, v_{q'}\right)/\tau}} \right). \tag{10}$$

Normalized Temperature-Scaled Cross Entropy is a loss function frequently employed in contrastive learning. It is a modified form of the cross-entropy loss, primarily utilized for learning meaningful representations in unlabeled data. While its initial applications were predominantly in image processing [12,56], it has also found a wide range of applications in the fields of IR [11,34]. In Equation (10), for any query intent $q$ in the batch, we assume that the original query representation of $q$ and its augmented query variation representation should be close to each other in the latent space, while it should be distant from the representations of other query intents within the input batch in the latent space. This alignment task enhances the PLM capability of natural language understanding but also contributes to the performance of the retrieval ranking task. Finally, we add the loss of the retrieval ranking and the query intent alignment task to conduct our contrastive ranking loss and train both Dense Retriever and Reranker:

$$Loss = \mathcal{L}_{BPR} + \alpha \cdot \mathcal{L}_{cl}, \tag{11}$$

where $\alpha$ is proposed to weight the training loss of both tasks.

## 4. Experiment

In this section, we will describe the details of the experimental implementation. First, we introduce the basic datasets used in the experiments and how to generate query variations using the GPT-3.5-turbo API. Next, we describe the evaluation metrics and baselines employed in the experiments. Finally, we present the experimental details during the training process.

### *4.1. Dataset*

In our experiment, we used two public datasets to evaluate the performance of the proposed approach:

- **WikiQA** [41]: the WikiQA corpus is a public dataset of question and sentence pairs collected and annotated for research on open-domain question answering. The dataset uses Bing query logs as the question source and includes 3047 questions and 29,258 sentences, where 1473 sentences were labeled answer sentences to their corresponding questions. Because the task is to find the answer for a given query and the label set is $\{0, 1\}$, we could regard it as an IR task that the IR model learns to retrieve the answer sentence from the candidate sentences when given a query.
- **ANTIQUE** [42]: ANTIQUE is an open field non-factoid question and answer dataset collected from community question and answer services. The collected data has a variety of categories, it contains 2626 open domain non-factor questions and 34,011 manual relevance annotations. These questions were asked by real users of the question and answer service: Yahoo! Answers and the relevance of all answers to each question is annotated by crowd-sourcing. Although candidate answers have four levels of relevance labels, we follow previous research [42,57], assuming that labels 3

and 4 are relevant labels, while labels 1 and 2 are irrelevant labels, which converts the four levels of relevance labels to two levels of relevance labels.

We obtained augmented query variation data on two datasets using the GPT-3.5-turbo API. The following is an example instruction:

> ***Paraphrase the following query and provide 5 different versions that convey the same meaning. Make sure each version is grammatically correct and clearly written.***
> ***Query: HOW AFRICAN AMERICANS WERE IMMIGRATED TO THE US***

After preprocessing the generated text, clean and high-quality query variations would be produced.

### *4.2. Evaluation Metrics*

Following the existing IR research, we selected MAP, MRR, NDCG@10, and P@10 to evaluate the retrieval ranking performance of our proposed method. Higher results for these metrics indicated better average retrieval performance. Moreover, the average retrieval performance was not the sole optimization target in our study. For different expressions of a query with the same intent, a robust retrieval ranking model should exhibit consistent retrieval performance with the original query. We assessed the model's robustness by measuring the decline in retrieval metrics under adversarial attacks, which was the average drop percentage (avg d.) and the worst drop percentage (worst d.). Lower results for these metrics indicated a more robust model.

### *4.3. Baselines*

To demonstrate the superiority of our proposed approach in terms of retrieval performance, we selected mainstream IR models as baselines, such as traditional BM25 and RM3 models, as well as neural network-based DRMM and KNRM. BM25 [58] was a probabilistic IR model that extends the Binary Independence Model (BIM) by incorporating term frequency and document length normalization. It has been proven to be effective in various information retrieval tasks and serves as a strong baseline for many modern approaches. RM3 (Relevance Model 3) [59] was a query expansion technique that employs pseudo-relevance feedback to improve retrieval performance. By estimating the relevance model from top-ranked documents and using it to expand the original query, RM3 was able to capture more relevant terms and provide better retrieval results. DRMM (Deep Relevance Matching Model) [60] was a neural network-based model that captures local interactions between query and document terms using histograms. It employed feed-forward neural networks to learn term-matching patterns and aggregated them to produce a final relevance score. KNRM (Kernelized Neural Ranking Model) [61] was another neural network-based model that integrates term-matching signals into a learning-to-rank framework using kernels. It mapped the semantic matching signals into a continuous space, allowing the model to learn fine-grained relevance patterns.

Additionally, the baselines included the original PLM-based Dense Retriever and Cross Encoder (Reranker) architectures. There are also some excellent studies in the community that focus on retrieval robustness. In this paper, we explored two other advanced robust retrieval methods, the typo-aware model [10] and Dual Contrastive Learning [11], to showcase and compare the robustness of different robust-enhanced retrieval methods. In the typo-aware method, the authors addressed the issue of decreased retrieval performance due to typos in queries, by proposing a simple spelling typo-aware training framework for both Dense Retriever and Reranker. In the Dual Contrastive Learning method, the authors analyzed the distribution of embeddings in the space and proposed an effective training paradigm that utilized contrastive learning to learn optimized representations for queries and documents, ensuring smoothness and uniformity in the embedding space. Note that Dual Contrastive Learning was only proposed for the DR architecture. To ensure fairness in the experiments, whenever a method involves using PLM as the encoder, we set the encoder to bert-base-uncased from the Hugging Face community.

### *4.4. Implementation Details*

We implemented our training framework based on both the Dense Retriever and Cross Encoder (Reranker) approaches. In the experiment, we designed an instruction to generate five different versions of certain query intent. We assumed that each variation generated by ChatGPT is uniformly distributed among the five samples. Therefore, during training, we randomly selected four variations as augmented samples, and the remaining one as an adversarial sample to test the robustness of the model trained with our method. In the data processing with positive and negative document pairings, we set the ratio of positive to negative samples for each query to be 1:4. For queries in the training set with fewer than four original negative samples, we globally randomly selected negative samples from the training set as a supplement. When implementing the Dense Retriever, the maximum query input sequence length was set to 64, and the maximum document input sequence length was set to 256. When implementing the Reranker, the combined maximum input sequence length for queries and documents was set to 256. Apart from the PLM-based IR models, various baseline models were implemented using Pyterrier [62], while neural network-based retrieval models were realized using OpenNIR [57]. In the stage of model training, we set the learning rate for PLM to $1 \times 10^{-5}$. and the learning rate for other neural network layers to $1 \times 10^{-3}$, and used AdamW [63] as the optimizer to train the model. We trained our models with four GeForce RTX 3090 GPUs each with 24 GB memory, and the inference program ran on the same GPUs. During the evaluation phase, we randomly selected an adversarial sample from the test set to attack the model, in order to assess the model's robustness.

## 5. Evaluation Result

### *5.1. Analysis of the Retrieval Effectiveness*

Tables 4 and 5 showed the retrieval effectiveness of IR models on the datasets. The results showed that PLM-based IR models demonstrate significant improvements in retrieval performance compared to traditional IR models and neural network-based IR models. In comparison to the original Dense Retriever (DR) and Reranker methods, our proposed method also exhibited superior performance. We believed that, under the condition of using the same ranking training method, the contrastive training method using language model-augmented query variations could fine-tune the language model to enhance the representation performance of query intent, reduce the risk of overfitting, and improve the generalizability of the model. Additionally, we also found that the retrieval performance of models based on Dense Retriever was generally inferior to that of the Reranker.

**Table 4.** Retrieval performance of each model on WikiQA. The best result is **bold-faced**.

| Model | MAP | MRR@10 | MRR | nDCG@10 | P@10 |
|---|---|---|---|---|---|
| BM25 | 0.569 | 0.575 | 0.601 | 0.659 | 0.110 |
| RM3 | 0.575 | 0.586 | 0.587 | 0.672 | 0.113 |
| DRMM | 0.612 | 0.617 | 0.620 | 0.698 | 0.113 |
| KNRM | 0.645 | 0.657 | 0.659 | 0.723 | 0.112 |
| DR | 0.781 | 0.792 | 0.793 | 0.833 | 0.115 |
| Reranker | 0.830 | 0.843 | 0.848 | 0.873 | 0.117 |
| Our (DR) | 0.785 | 0.801 | 0.801 | 0.841 | 0.117 |
| Our (Reranker) | **0.845** | **0.862** | **0.862** | **0.888** | **0.119** |

**Table 5.** Retrieval performance of each model on ANTIQUE. The best result is **bold-faced**.

| Model | MAP | MRR@10 | MRR | nDCG@10 | P@10 |
|---|---|---|---|---|---|
| BM25 | 0.192 | 0.506 | 0.509 | 0.510 | 0.238 |
| RM3 | 0.176 | 0.433 | 0.437 | 0.488 | 0.228 |
| DRMM | 0.187 | 0.515 | 0.530 | 0.404 | 0.234 |
| KNRM | 0.203 | 0.574 | 0.579 | 0.443 | 0.255 |
| DR | 0.438 | 0.677 | 0.679 | 0.578 | 0.317 |
| Reranker | 0.511 | 0.747 | 0.752 | 0.654 | 0.363 |
| Our (DR) | 0.438 | 0.677 | 0.681 | 0.558 | 0.310 |
| Our (Reranker) | **0.529** | **0.770** | **0.771** | **0.659** | **0.367** |

### *5.2. Analysis of the Retrieval Robustness*

To comprehensively compare the robustness of different IR models, we designed experiments that calculate avg d. and worst d. for different models and training methods when using query variations for retrieval.

#### 5.2.1. Comparison Analysis on Retrieval Robustness of Different IR Models

The results shown in Figure 3 indicated that the method proposed in this study performed the best overall in terms of various metrics, both in avg d. and worst d. The decline in performance for the proposed method was the lowest among all models in almost all experiments, demonstrating a better resilience to adversarial attack compared to the original BERT-based Reranker model. In terms of the defensive ability against query attacks, the performance of traditional IR models and neural network-based IR models had their own advantages and disadvantages, which was consistent with the conclusions of the previous study [7]. Moreover, we noticed that the robustness of the models varies across different datasets: on WikiQA, the robustness of traditional IR models was superior to that of PLM-based IR models. Although the proposed method in this study could improve the robustness of the original PLM-based IR models, it still did not show a significant advantage compared to traditional models; on ANTIQUE, PLM-based IR models outperformed traditional IR models completely. We believed that this result was related to the number and quality of candidate document lists in the dataset, as well as the levels and quantities of relevance annotations. In comparison to ANTIQUE, WikiQA had a smaller number of candidate documents in the test set and relatively simpler query-document matching tasks, resulting in a less pronounced gap. On the other hand, ANTIQUE had more candidates, higher quality relevance annotations, and some more difficult query tasks, leading to a larger gap in the results of different models.

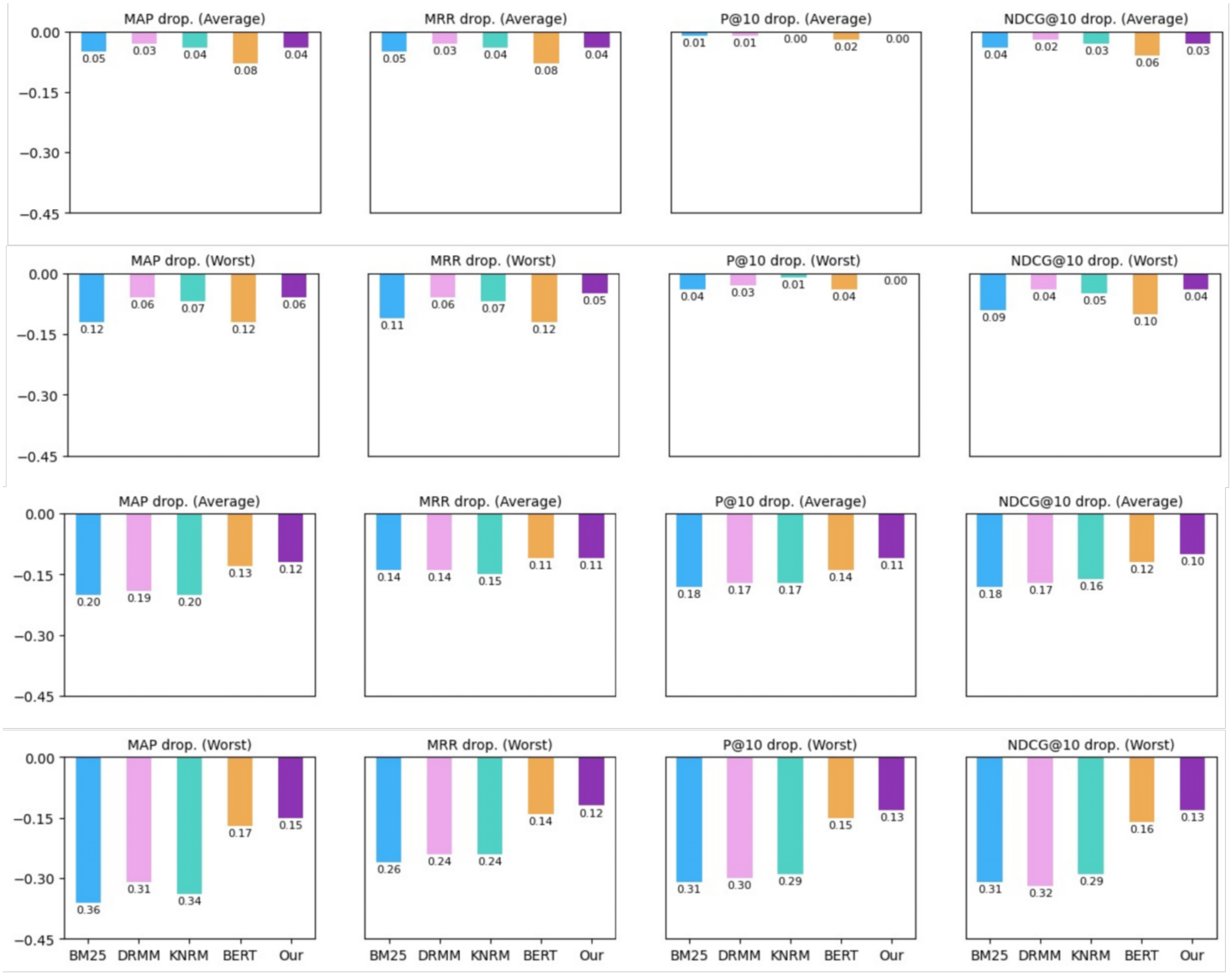

**Figure 3.** The avg d. and worst d. of MAP, MRR, P@10, and NDCG@10 of different IR models. We selected a subset of traditional retrieval models from the baselines, as well as PLM-based IR models, represented by BERT. Both BERT and our proposed method employed Cross Encoder (Reranker). The data in the top two rows represent the results for WikiQA, while the data in the bottom two rows represent the results for ANTIQUE. The values shown in the subfigures represent the absolute values of the actual results. A higher value indicates poorer robustness. Each color signifies the robustness of the corresponding model under different datasets and evaluation metrics.

### 5.2.2. Comparison Analysis on Retrieval Robustness of Different Robustness Training Approaches

The results shown in Tables 6 and 7 reflected the robustness of the models under different robustness training methods when dealing with adversarial attacks. We calculated the evaluation metrics for the IR models and demonstrated the extent of the decline in their performance when subjected to attacks. The Dual Contrastive method on the DR model achieved the best retrieval performance, but its defensive ability against query attacks did not show a significant improvement compared to the original model. We believe that the improvement was mainly attributed to the optimization of query and document representations in the latent space during model training, which further enhanced the PLM's ability to represent documents. Although slightly inferior to Dual Contrastive in retrieval performance, our proposed approach could significantly improve the robustness of the model. The typo-aware method, on the other hand, exhibited excellent robustness when facing adversarial attacks despite a slight decline in retrieval performance compared to the original method. Additionally, the performance of different training methods varied across different datasets. On WikiQA, the typo-aware method demonstrated the best robustness for the DR model, whereas our method exhibited the best robustness in all other experiments. Considering both retrieval performance and robustness as training objectives, the method proposed in this study achieved better robustness while ensuring nearly optimal retrieval performance.

**Table 6.** Results of different robustness training approaches on WikiQA. The best accuracy is **bold-faced**, and the second-best accuracy is underlined.

| Model Type | Method | MAP | MRR | nDCG@10 | P@10 |
|---|---|---|---|---|---|
| DR | Original | 0.781(4.1%/12.3%) | 0.793(4.1%/11.8%) | 0.833(2.9%/8.6%) | 0.115(**0.0%/0.0%**) |
| | Typo-aware | 0.779(**2.8%/4.6%**) | 0.791(**2.7%/4.1%**) | 0.832(**1.9%/3.0%**) | 0.116(0.0%/0.4%) |
| | Dual contrastive | **0.789**(3.9%/7.9%) | **0.802**(3.9%/6.7%) | **0.843**(2.8%/5.5%) | **0.117**(0.3%/0.7%) |
| | Our | 0.785(3.6%/5.9%) | 0.801(3.7%/5.4%) | 0.841(2.7%/4.0%) | **0.117**(0.1%/0.3%) |
| Reranker | Original | 0.830(7.8%/12.2%) | 0.848(8.3%/12.3%) | 0.873(6.2%/9.7%) | 0.117(1.7%/3.6%) |
| | Typo-aware | 0.819(4.9%/11.5%) | 0.832(4.9%/10.6%) | 0.866(3.7%/8.3%) | 0.117(0.4%/1.1%) |
| | Our | **0.845(3.7%/7.2%)** | **0.862(3.6%/6.7%)** | **0.888(2.7%/5.0%)** | **0.119(0.1%/0.7%)** |

**Table 7.** Results of different robustness training approaches on ANTIQUE. The best accuracy is **bold-faced**, and the second-best accuracy is underlined.

| Model Type | Method | MAP | MRR | nDCG@10 | P@10 |
|---|---|---|---|---|---|
| DR | Original | **0.438**(17.0%/24.1%) | 0.679(15.3%/23.6%) | **0.578**(15.7%/22.3%) | **0.316**(18.1%/24.7%) |
| | Typo-aware | 0.427(12.4%/17.5%) | 0.671(11.0%/16.5%) | 0.565(11.4%/16.3%) | 0.312(13.3%/18.0%) |
| | Dual contrastive | 0.437(17.4%/24.1%) | **0.689**(15.7%/22.7%) | **0.578**(15.8%/21.5%) | 0.314(17.2%/22.9%) |
| | Our | **0.438(10.2%/12.7%)** | 0.682(**8.9%/12.1%**) | 0.558(**9.1%/11.8%**) | 0.313(**9.3%/13.1%**) |
| Reranker | Original | 0.511(13.0%/16.5%) | 0.752(10.8%/14.5%) | 0.654(13.7%/15.1%) | 0.363(12.1%/16.2%) |
| | Typo-aware | 0.527(13.1%/16.6%) | **0.776**(10.9%/13.0%) | 0.652(11.2%/14.2%) | 0.360(11.8%/14.7%) |
| | Our | **0.529(11.9%/15.2%)** | 0.771(**10.6%/12.9%**) | **0.659(10.5%/13.7%)** | **0.368(10.2%/13.2%)** |

### 5.3. Case Study

To verify the actual improvement in robustness of the models, we used real data from the ANTIQUE dataset as the subject for analysis. Following the idea shown in Table 1, we selected existing adversarial texts and compared the performance differences of PLM-based IR models trained using different training methods.

The results presented in Table 8 showed that users probably make spelling errors during the actual retrieval process, including omissions of some letters and swapping letter positions. Based on the original retrieval metrics, it could be observed that these potential adversarial attacks did pose challenges to the retrieval performance of IR models, especially when the target of the attack was the keyword in the query. Traditional models based on word-matching would be highly likely to fail in matching, while PLM-based IR models may also experience a decline in retrieval performance due to the reduced representation performance after being attacked. For example, in the first case in the table with the misspelling *"rabit"*, the MAP and MRR values of the original model were very low, and the relevant document was ranked seventh in terms of MRR. However, after being trained with our proposed robustness framework, the relevant document was ranked third. Of course, it was also possible for common spelling errors to occur simultaneously in both documents and queries, such as in the second case, where "potatoes" was often misspelled as *"potatos"*. Although the MRR of the PLM-based IR model was equal to 1 under both training methods, our proposed approach could also improve the other MAP indicator. This suggested that our proposed method could learn the semantic representation of queries during training and thus improved the ranking positions of other marginally relevant documents in the retrieval process. Overall, the robustness training method proposed in this study enhanced the robustness of pre-trained models.

**Table 8.** Case studies. Three cases are selected from ANTIQUE to qualitatively analyze the ranking effectiveness of our proposed method under a query attack.

| Real Query and Its Relevant Document | Original (MAP/MRR) | Our (MAP/MRR) |
|---|---|---|
| **Q**: How can I keep my rabit indoors? #4473331<br>**D**: just bring the cage indoors and keep it clean. . . Oh, and don't confuse the cocoa puffs with the rabbit droppings!! | 0.050/0.142 | 0.245/0.333 |
| **Q**: why do my baked potatos never taste near as good as when I get them in a nice restaurant? #402514<br>**D**: Due to the cooking time a lot of restaurants bake potatos twice. Wrap them in foil, with some course sea salt for the first bake. Allow them to cool, unwrap them and bake again. | 0.295/1.000 | 0.612/1.000 |
| **Q**: how do I go about getting copies of letters of commendation prsented to me from city of san diego? #100653<br>**D**: Use a copy machine. Or scan the letter into your computer and print it out in color mode to pick up any logos on the original letterhead. | 0.159/0.250 | 0.336/0.333 |

## 6. Discussion

### *6.1. Limitations and Future Work*

This study follows the prevalent modeling approach in IR, which involves fine-tuning pre-trained models such as BERT to achieve semantic matching capabilities for retrieval tasks. With the remarkable general capabilities demonstrated by LLMs in the NLP arena, it becomes pertinent to ponder on how to further harness LLMs for enhanced performance in information retrieval tasks. Firstly, this paper predominantly focuses on the fine-tuning of existing pre-trained models to bolster the encoder's robustness to queries while ensuring retrieval performance. As per the research by Wang et al. [38], LLMs exhibit a distinct advantage in robustness compared to smaller-scale models such as BERT, symbolizing a tremendous leap in linguistic comprehension. Consequently, enhancing the robustness of pre-trained models to fundamentally address challenges associated with downstream tasks will be a more foundational and exhilarating endeavor. Secondly, we observed that the robustness disparity across various retrieval models on WikiQA is significantly less than on ANTIQUE. We postulate that this might stem from differences in task difficulty across datasets. A more profound evaluation is therefore required to validate this hypothesis to investigate the factors that influence the robustness performance of IR models. Lastly, although this paper employs LLM-generated query variations that maintain the original query's intent and exhibit a richer diversity in expression, they do not perfectly align with the real distribution of user queries. Whether LLMs can be utilized to generate queries that truly reflect human-generated statements remains a particularly intriguing direction for exploration.

### *6.2. Ethical Considerations*

This paper introduces a method to enhance the robustness of retrieval models using data generated by LLMs. Inherently, the method does not entail any ethical considerations. However, it is essential to note that the outcomes generated by LLMs possess a certain degree of randomness. Taking ethical considerations into account, it is recommended that when employing the method proposed in this study, one should ensure that the results produced by large language models do not contain any unethical content.

### *6.3. Real-World Applicability*

The method proposed in this paper can be flexibly adapted to the training processes of all IR models based on PLMs, whether they are built upon the Dense Encoder or Reranker architectures. It is worth noting that while our method indeed presents vast long-term application prospects and reliable performance, these outcomes are contingent upon dependable training data: specifically, high-quality LLM-generated query variations.

## 7. Conclusions

In this paper, we propose a training method for enhancing the defensive ability against query attacks in PLM-based IR models, which can be applied to both the mainstream retrieval structures: Dense Retriever and Reranker. To obtain query expansions with the same intent, we first employ ChatGPT to generate adversarial samples. Compared to rule-based and seq2seq model-based expansion methods, we find that the adversarial samples generated by LLMs exhibit greater diversity and higher quality, and can automatically correct possible spelling and grammatical errors in the original query. Secondly, we propose using contrastive learning to improve the robustness of IR models. We employ ranking contrastive learning to train the model's ranking capabilities while optimizing the representation of similar semantic texts in latent space for different query expressions under the same intent. This results in more uniform and distinguishable query embeddings in latent space. Ultimately, we merge the two contrastive learning losses through multi-task training to achieve robustness training for IR models. This study conducts experiments on two publicly available datasets, WikiQA and ANTIQUE. The experimental results indicate that current state-of-the-art IR models, such as PLM-based IR models, still face challenges from adversarial attacks. Our proposed method could significantly enhance the robustness of IR models while maintaining their retrieval performance. In the experimental results, we also discover that our proposed method cannot maintain the best performance across all models, which is a limitation of the method presented in this study.

**Author Contributions:** Conceptualization, Z.P.; Data curation, K.F. and R.L.; Formal analysis, Z.P.; Funding acquisition, D.L.; Investigation, R.L.; Methodology, Z.P.; Project administration, D.L.; Software, Z.P. and K.F.; Supervision, D.L.; Validation, Z.P. and K.F.; Writing—original draft, Z.P.; Writing—review & editing, K.F., R.L. and D.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the National Natural Science Foundation of China (NSFC) under grant no. 72074231.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** To verify the effectiveness of the proposed model, we utilized two datasets - WikiQA and ANTIQUE. WikiQA is a public dataset provided by Microsoft: http://aka.ms/WikiQA (accessed on 30 July 2023). ANTIQUE is a public dataset which can be downloaded on https://ciir.cs.umass.edu/downloads/Antique/ (accessed on 30 July 2023) for research purposes.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
2. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.
3. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
4. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 5485–5551.
5. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744.
6. Fan, Y.; Xie, X.; Cai, Y.; Chen, J.; Ma, X.; Li, X.; Zhang, R.; Guo, J. Pre-training Methods in Information Retrieval. *Found. Trends Inf. Retr.* **2022**, *16*, 178–317. https://doi.org/10.1561/1500000100.
7. Wu, C.; Zhang, R.; Guo, J.; Fan, Y.; Cheng, X. Are Neural Ranking Models Robust? *ACM Trans. Inf. Syst.* **2022**, *41* , 1–36. https://doi.org/10.1145/3534928.

8. Sidiropoulos, G.; Kanoulas, E. Analysing the Robustness of Dual Encoders for Dense Retrieval Against Misspellings. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, 11–15 July 2022; pp. 2132–2136. https://doi.org/10.1145/3477495.3531818.
9. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
10. Zhuang, S.; Zuccon, G. Dealing with Typos for BERT-based Passage Retrieval and Ranking. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 2836–2842. https://doi.org/10.18653/v1/2021.emnlp-main.225.
11. Li, Y.; Liu, Z.; Xiong, C.; Liu, Z. More Robust Dense Retrieval with Contrastive Dual Learning. In Proceedings of the ICTIR '21: The 2021 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual, 11 July 2021; pp. 287–296. https://doi.org/10.1145/3471158.3472245.
12. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria, 12–18 July 2020.
13. Zhang, W.E.; Sheng, Q.Z.; Alhazmi, A.; Li, C. Adversarial Attacks on Deep-Learning Models in Natural Language Processing: A Survey. *ACM Trans. Intell. Syst. Technol.* **2020**, *11*, 1–41. https://doi.org/10.1145/3374217.
14. Morris, J.X.; Lifland, E.; Yoo, J.Y.; Grigsby, J.; Jin, D.; Qi, Y. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. *arXiv* **2020**, arXiv:2005.05909.
15. Li, J.; Ji, S.; Du, T.; Li, B.; Wang, T. TextBugger: Generating Adversarial Text Against Real-world Applications. In Proceedings of the Proceedings Network and Distributed System Security Symposium, San Diego, CA, USA, 24–27 February 2019. https://doi.org/10.14722/ndss.2019.23138.
16. Gao, L.; Dai, Z.; Callan, J. Rethink Training of BERT Rerankers in Multi-Stage Retrieval Pipeline. In Proceedings of the Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, 1 April–28 March 2021; pp. 280–286. https://doi.org/10.1007/978-3-030-72240-1_26.
17. Gao, L.; Dai, Z.; Chen, T.; Fan, Z.; Van Durme, B.; Callan, J. Complement Lexical Retrieval Model with Semantic Residual Embeddings. In *Proceedings of the Advances in Information Retrieval*; Hiemstra, D., Moens, M.F., Mothe, J., Perego, R., Potthast, M., Sebastiani, F., Eds.; Springer: Cham, Switzerland, 2021; pp. 146–160.
18. Crammer, K.; Singer, Y. Pranking with Ranking. In *Proceedings of the Advances in Neural Information Processing Systems*; Dietterich, T., Becker, S., Ghahramani, Z., Eds.; MIT Press: Cambridge, MA, USA, 2001; Volume 14.
19. Shashua, A.; Levin, A. Ranking with Large Margin Principle: Two Approaches. In *Proceedings of the Advances in Neural Information Processing Systems*; Becker, S., Thrun, S., Obermayer, K., Eds.; MIT Press: Cambridge, MA, USA, 2002; Volume 15.
20. Burges, C.; Shaked, T.; Renshaw, E.; Lazier, A.; Deeds, M.; Hamilton, N.; Hullender, G. Learning to Rank Using Gradient Descent. In Proceedings of the 22nd International Conference on Machine Learning, New York, NY, USA, 7–11 August 2005; pp. 89–96. https://doi.org/10.1145/1102351.1102363.
21. Burges, C.; Ragno, R.; Le, Q. Learning to Rank with Nonsmooth Cost Functions. In *Proceedings of the Advances in Neural Information Processing Systems*; Schölkopf, B., Platt, J., Hoffman, T., Eds.; MIT Press: Cambridge, MA, USA, 2006; Volume 19.
22. Joachims, T. Optimizing Search Engines Using Clickthrough Data. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 23–26 July 2002; pp. 133–142. https://doi.org/10.1145/775047.775067.
23. Cao, Z.; Qin, T.; Liu, T.Y.; Tsai, M.F.; Li, H. Learning to Rank: From Pairwise Approach to Listwise Approach. In Proceedings of the 24th International Conference on Machine Learning, New York, NY, USA, 20–24 June 2007; pp. 129–136. https://doi.org/10.1145/1273496.1273513.
24. Xu, J.; Li, H. AdaRank: A Boosting Algorithm for Information Retrieval. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, 23–27 July 2007; pp. 391–398. https://doi.org/10.1145/1277741.1277809.
25. Wu, Q.; Burges, C.J.C.; Svore, K.M.; Gao, J. Adapting boosting for information retrieval measures. *Inf. Retr.* **2010**, *13*, 254–270. https://doi.org/10.1007/s10791-009-9112-1.
26. Arslan, A.; Dinçer, B.T. A selective approach to index term weighting for robust information retrieval based on the frequency distributions of query terms. *Inf. Retr. J.* **2019**, *22*, 543–569. https://doi.org/10.1007/s10791-018-9347-9.
27. Gangi Reddy, R.; Yadav, V.; Arafat Sultan, M.; Franz, M.; Castelli, V.; Ji, H.; Sil, A. Towards Robust Neural Retrieval Models with Synthetic Pre-Training. *arXiv* **2021**, arXiv:2104.07800. https://doi.org/10.48550/arXiv.2104.07800.
28. Prakash, P.; Killingback, J.; Zamani, H. Learning Robust Dense Retrieval Models from Incomplete Relevance Labels. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, 11–15 July 2021; pp. 1728–1732. https://doi.org/10.1145/3404835.3463106.
29. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R.B. Momentum Contrast for Unsupervised Visual Representation Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9726–9735.
30. Zhang, Y.; He, R.; Liu, Z.; Lim, K.; Bing, L. An Unsupervised Sentence Embedding Method by Mutual Information Maximization In Proceedings of the Conference on Empirical Methods in Natural Language Processing—EMNLP 2020, Online, 16–20 November 2015 http://dx.doi.org/10.18653/v1/2020.emnlp-main.124.

31. Cheng, P.; Hao, W.; Yuan, S.; Si, S.; Carin, L. FairFil: Contrastive Neural Debiasing Method for Pretrained Text Encoders. In Proceedings of the 9th International Conference on Learning Representations—ICLR 2021, Virtual Event, Austria, 3–7 May 2021.
32. Qu, Y.; Shen, D.; Shen, Y.; Sajeev, S.; Chen, W.; Han, J. CoDA: Contrast-enhanced and Diversity-promoting Data Augmentation for Natural Language Understanding. In Proceedings of the 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, 3–7 May 2021.
33. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, 3–7 November 2019*; Inui, K., Jiang, J., Ng, V., Wan, X., Eds.; Association for Computational Linguistics: Toronto, ON, Canada, 2019; pp. 3980–3990. https://doi.org/10.18653/v1/D19-1410.
34. Izacard, G.; Caron, M.; Hosseini, L.; Riedel, S.; Bojanowski, P.; Joulin, A.; Grave, E. Towards Unsupervised Dense Information Retrieval with Contrastive Learning. *arXiv* **2021**, arXiv:2112.09118.
35. Deng, Y.; Zhang, W.; Lam, W. Learning to Rank Question Answer Pairs with Bilateral Contrastive Data Augmentation. In *Proceedings of the Seventh Workshop on Noisy User-Generated Text, W-NUT 2021, Online, 11 November 2021*; Xu, W., Ritter, A., Baldwin, T., Rahimi, A., Eds.; Association for Computational Linguistics: Toronto, ON, Canada, 2021; pp. 175–181. https://doi.org/10.18653/v1/2021.wnut-1.20.
36. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, 5–10 July 2020*; Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R., Eds.; Association for Computational Linguistics: Toronto, ON, Canada, 2020; pp. 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703.
37. OpenAI. Introducing ChatGPT. Available online: https://openai.com/blog/chatgpt (accessed on 16 April 2023).
38. Wang, J.; Hu, X.; Hou, W.; Chen, H.; Zheng, R.; Wang, Y.; Yang, L.; Huang, H.; Ye, W.; Geng, X.; et al. On the Robustness of ChatGPT: An Adversarial and Out-of-distribution Perspective. *arXiv* **2023**, arXiv:2302.12095.
39. Penha, G.; Câmara, A.; Hauff, C. Evaluating the Robustness of Retrieval Pipelines with Query Variation Generators. In Proceedings of the Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, 10–14 April 2022; pp. 397–412. https://doi.org/10.1007/978-3-030-99736-6_27.
40. Fan, A.; Bhosale, S.; Schwenk, H.; Ma, Z.; El-Kishky, A.; Goyal, S.; Baines, M.; Celebi, O.; Wenzek, G.; Chaudhary, V.; et al. Beyond English-Centric Multilingual Machine Translation. *J. Mach. Learn. Res.* **2021**, *22*, 107:1–107:48.
41. Yang, Y.; Yih, W.; Meek, C. WikiQA: A Challenge Dataset for Open-Domain Question Answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing—EMNLP 2015, Lisbon, Portugal, 17–21 September 2015*; Màrquez, L., Callison-Burch, C., Su, J., Pighin, D., Marton, Y., Eds.; The Association for Computational Linguistics: Toronto, ON, Canada, 2015; pp. 2013–2018. https://doi.org/10.18653/v1/d15-1237.
42. Hashemi, H.; Aliannejadi, M.; Zamani, H.; Croft, W.B. ANTIQUE: A Non-factoid Question Answering Benchmark. In *Proceedings of the Advances in Information Retrieval—42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, 14–17 April 2020*; Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F., Eds.; Springer: Berlin/Heidelberg, Germany, 2020; Volume 12036, pp. 166–173. https://doi.org/10.1007/978-3-030-45442-5_21.
43. Muennighoff, N.; Tazi, N.; Magne, L.; Reimers, N. MTEB: Massive Text Embedding Benchmark. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Dubrovnik, Croatia, 2–4 May 2023; pp. 2014–2037.
44. BAAI. Bge-Large-En. 2023 Available online: https://huggingface.co/BAAI/bge-large-en (accessed on 27 August 2023).
45. BAAI. Bge-Base-En. 2023 Available online: https://huggingface.co/BAAI/bge-base-en (accessed on 27 August 2023).
46. Su, H.; Shi, W.; Kasai, J.; Wang, Y.; Hu, Y.; Ostendorf, M.; Yih, W.t.; Smith, N.A.; Zettlemoyer, L.; Yu, T. One Embedder, Any Task: Instruction-Finetuned Text Embeddings. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, Toronto, ON, Canada, 9–14 July 2023; pp. 1102–1121. https://doi.org/10.18653/v1/2023.findings-acl.71.
47. Li, Z.; Zhang, X.; Zhang, Y.; Long, D.; Xie, P.; Zhang, M. Towards General Text Embeddings with Multi-stage Contrastive Learning. *arXiv Prepr.* **2023** arXiv:2308.03281.
48. Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; Yih, W.t. Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550.
49. Xiong, W.; Li, X.; Iyer, S.; Du, J.; Lewis, P.; Wang, W.Y.; Mehdad, Y.; Yih, S.; Riedel, S.; Kiela, D.; et al. Answering Complex Open-Domain Questions with Multi-Hop Dense Retrieval. In Proceedings of the International Conference on Learning Representations, Virtual, 3–7 May 2021. .
50. Rendle, S.; Freudenthaler, C.; Gantner, Z.; Schmidt-Thieme, L. BPR: Bayesian Personalized Ranking from Implicit Feedback. In Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, Arlington, VA, USA, 18–21 June 2009; pp. 452–461.
51. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder—Decoder for Statistical Machine Translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 26–28 October 2014; pp. 1724–1734. https://doi.org/10.3115/v1/D14-1179.

52. Ji, S.; Yun, H.; Yanardag, P.; Matsushima, S.; Vishwanathan, S.V.N. WordRank: Learning Word Embeddings via Robust Ranking. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 658–668. https://doi.org/10.18653/v1/D16-1063.
53. Otegi, A.; San Vicente, I.; Saralegi, X.; Peñas, A.; Lozano, B.; Agirre, E. Information retrieval and question answering: A case study on COVID-19 scientific literature. *Knowl. Based Syst.* **2022**, *240*, 108072. https://doi.org/https://doi.org/10.1016/j.knosys.2021.108072.
54. Ulukus, S.; Avestimehr, S.; Gastpar, M.; Jafar, S.A.; Tandon, R.; Tian, C. Private Retrieval, Computing, and Learning: Recent Progress and Future Challenges. *IEEE J. Sel. Areas Commun.* **2022**, *40*, 729–748. https://doi.org/10.1109/JSAC.2022.3142358.
55. Wang, H.; Jia, Y.; Wang, H. Interactive Information Retrieval with Bandit Feedback. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, 11–15 July 2021; pp. 2658–2661. https://doi.org/10.1145/3404835.3462810.
56. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. *arXiv* **2020**, arXiv:1911.05722.
57. MacAvaney, S. OpenNIR: A Complete Neural Ad-Hoc Ranking Pipeline. In Proceedings of the 13th International Conference on Web Search and Data Mining, New York, NY, USA, 2020; pp. 845–848. https://doi.org/10.1145/3336191.3371864.
58. Robertson, S.E.; Walker, S. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin Ireland, 3–6 July 1994; pp. 232–241.
59. Lavrenko, V.; Croft, W.B. Relevance Based Language Models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*; Association for Computing Machinery: New York, NY, USA, 2001; pp. 120–127. https://doi.org/10.1145/383952.383972.
60. Guo, J.; Fan, Y.; Ai, Q.; Croft, W.B. A Deep Relevance Matching Model for Ad-Hoc Retrieval. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, Indianapolis, IN, USA, 24–18 October 2016; pp. 55–64. https://doi.org/10.1145/2983323.2983769.
61. Xiong, C.; Dai, Z.; Callan, J.; Liu, Z.; Power, R. End-to-End Neural Ad-hoc Ranking with Kernel Pooling. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Tokyo, Japan, 7–11 August 2017. https://doi.org/10.1145/3077136.3080809.
62. Macdonald, C.; Tonellotto, N. Declarative Experimentation in Information Retrieval Using PyTerrier. In Proceedings of the ACM SIGIR on International Conference on Theory of Information Retrieval, Virtual, 14–17 September 2020; pp. 161–168. https://doi.org/10.1145/3409256.3409829.
63. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.