Robust Dense Retrieval with Conterfactual Learning

Anonymous ACL submission

Abstract

Efficiently retrieving a concise set of candidates from an extensive document corpus remains a pivotal challenge in Information Retrieval (IR). Dense retrieval models have gained prominence for their superior performance. However, criticisms surrounding their vulnerability to adversarial attacks necessitate further exploration. In response to these challenges, we advocate for enhancing the sensitivity of dense retrieval models to fine-grained relevance matching signals. Motivated by causality and counterfactual analysis, we propose a series of counterfactual regularization methods. These methods aim to improve the robustness and sensitivity of dense retrieval models concerning fine-grained relevance matching. We first introduce a counterfactual passage extraction method, identifying the key passages that can influence relevance. Subsequent unsupervised learning tasks, based on these counterfactual passages, serve to regularize the model's learning process to improve the robustness and sensitivity. Experimental results showcase the efficacy of our approach. Our method demonstrates the ability to learn key passages without reliance on the passage-level relevance annotations. Moreover, the regularized dense retrieval models exhibit heightened robustness against adversarial attacks, surpassing the state-of-the-art anti-attack methods. Our work contributes to bridging the vulnerability gaps in dense retrieval models, paving the way for their improved reliability and applicability in IR systems.

1 Introduction

How to efficiently retrieve a concise set of candidates from an extensive pool of documents is a fundamental challenge in Information Retrieval (IR). In recent years, neural retrieval models have attracted considerable attention for their superior performance. Dense retrieval models, in particular, have not only achieved state-of-the-art retrieval results but have also demonstrated comparable efficiency to traditional systems based on term matching models and inverted indexes (Robertson et al., 2009; Ramos et al., 2003).

However, the susceptibility of modern retrieval models to adversarial attacks poses a critical limitation. Deep learning-based models, including dense retrieval models, are vulnerable to the injection of input "noise," leading to significant variations in the treatment of documents with imperceptible differences to humans within the retrieval system. Previous studies (Liu et al., 2023a,b) have illustrated how replacing words with specific synonyms can deceive dense retrieval models, manipulating the ranking position of a target document regardless of its actual relevance to the query. Guided by these findings, we propose a training task to aid dense retrieval models in learning the distinction between positive documents and adversarial documents, thereby enhancing the models' robustness.

This paper posits that a key aspect of mitigating vulnerability is to augment the sensitivity of dense retrieval models to fine-grained relevance matching signals. The sensitivity of a particular input data segment generally reflects its significance within the neural network model. Ideally, a retrieval model should adeptly capture both the overall relevance of a document and the role each part of the document plays in fulfilling the information need of a given query. Therefore, we argue that a retrieval model demonstrates sensitivity to fine-grained relevance matching if its ranking scores for a specific query-document pair exhibit: (1) high variances when the key passages of the document that satisfies the query need are modified, and (2) low variances when the irrelevant passages of the document with respect to the current query are changed. The former indicates that the model should be sensitive to passages in a document that potentially determine its relevance to the query, while the later indicates that the model should be robust and stable with respect to changes on passages that are unimportant to the query. This approach makes the model more resistant to adversarial attacks, as any attacks on terms or passages that don't determine the relevance of a document would not lead to significant changes in model outputs.

To address this limitation, we propose a counterfactual passage extraction method based on the Shapley value (Winter, 2002). The objective is to identify the key passage of each document, termed the counterfactual passage, determining its relevance to a given query. Subsequently, we introduce several unsupervised learning tasks based on these counterfactual passages to regulate the learning

process of dense retrieval models. Experimental results demonstrate that our method and model can learn to extract key passages influencing the relevance between a document and a query without requiring passage-level relevance annotations. Additionally, the regularized dense retrieval models exhibit robustness against adversarial attacks, surpassing even state-of-the-art anti-attack methods designed specifically based on attack properties.

2 Related Work

2.1 Dense Retieval

With the development of the deep neural network, pretrained language models, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), have been widely used in both representation-based and interaction-based Neu-IR models. Much work (Khattab and Zaharia, 2020; Lewis et al., 2020) proposes a latent query–document interaction paradigm to improve the retrieval effectiveness, many others focus on learning better representations with contrastive training. These work propose various negative sampling methods during contrastive training, including in-batch negatives, BM25 retrieved negatives and randomly selected negatives (Lewis et al., 2020; Xiong et al., 2021a,b).

2.2 Counterfactual Learning

The counterfactual inference has been applied to representation learning to obtain fair representations (Kusner et al., 2017) in various domains such as image classification (Goyal et al., 2019) and vision-language tasks (Liang et al., 2020; Niu et al., 2021). The key idea behind counterfactual learning is to train a model that is invariant to specific aspects of the input data (Johansson et al., 2016; Kusner et al., 2017).

Existing work on counterfactual learning for natural language expands upon this idea and aims to learn robust representations of text data by capturing causal features while mitigating spurious correlations (Choi et al., 2020; Liang et al., 2020; Tian et al., 2022). However, few studies have delved into the effect of counterfactual learning on retrieval tasks.

2.3 Robustness of Dense Retrievers

In recent years, model robustness has attracted attention in various fields (Wang et al., 2022; Liang et al., 2024; Wang et al., 2023; Liang et al., 2023), including IR (Liu et al., 2023a,b). In the context of robustness, adversarial attacks aim to discover human-imperceptible perturbations that can deceive neural networks (Szegedy et al., 2014). Wu et al. (Wu et al., 2023a) introduced the WSRA method of attacking black-box NRMs using word substitution. This study revealed the serious vulnerability of NRMs to synonym substitution perturbations. As a result, subsequent explorations of attack against NRMs have emerged (Liu et al., 2023d; Chen et al., 2023), inspired by this pioneering work.

In response to adversarial attacks, research has proposed various defense strategies to enhance adversarial robustness. These can be generally classified into certified defenses and empirical defenses. Certified defenses aim for theoretical robustness against specific adversarial perturbations (Raghunathan et al., 2018). For instance, (Wu et al., 2022) introduced a certified defense method that ensures the top-K robustness of NRMs via randomized smoothing.

3 Methodology

3.1 Preliminary

Given a query q and a document collection $D = \{d_1, \ldots, d_n\}$, dense retrievers calculate the relevance score $f(q, d)$ based on the dense representations of the query and document. In particular, the representations of a query and a document is denoted as e_q and e_d, respectively. Then the similarity score, i.e. $f(q, d)$, of query q and document d can be calculated with their dense representations:

$$f(q, d) = sim(e_q, e_d), \tag{1}$$

where $sim(\cdot)$ is the similarity function, which is used to estimate the relevance between two embeddings. The dot product and cosine similarity is usually used as the similarity function.

To obtain the relevance score $f(q, d)$, the dense retrievers are typically trained using the triplet training samples, consisted of the query q and its positive (relevant) document d^+ and negative ducoment d^-. The dense retrieval models are optimised by minimising the following loss function:

$$L = \sum_q l(q, d^+, D^-), \tag{2}$$

where D^- is the collection of the negative samples for the query q, and $l(q, d^+, D^-)$ is the contrastive

training loss function, defined as following:

$$l(q, d^+, D^-) = -log \frac{e^{f(q,d^+)}}{e^{f(q,d^+)} + \sum_{d^- \in D^-} e^{f(q,d^-)}}. \quad (3)$$

However, this loss function only considers the contrastive learning between positive and negative examples to enhance retrieval effectiveness, neglecting the model's robustness.

3.2 Counterfactual Learning Method

Extracting key passages from documents is a crucial prerequisite for the sensitivity of the model. To accomplish this task, we employed a model-agnostic method for key passage extraction. On one hand, model-agnostic methods have already provided a certain level of interpretability to the model. On the other hand, they can supply appropriate data for constructing counterfactual documents in subsequent pretraining tasks.

For a given query q and document d, where the document d can be lengthy and thus consists of different passages $P_d = p_1, \ldots, p_n$, each passage may convey varying amounts of information. Different passages within a long document can exhibit different level of relevance to a query, thus possessing varying importance when measuring the relevance score of a query. However, in existing dense retrieval models, determining the importance of the passages within a document in relation to a search query is challenging.

Counterfactual learning has been widely employed to enhance the explainability of deep learning models. Some efforts have been made to implement counterfactual analysis in Information Retrieval (IR) (Song et al., 2023). However, effective implementation of counterfactual analysis for dense retrieval models still lacks comprehensive research. Therefore, in this work, we propose a counterfactual learning method specifically designed for dense retrieval. In particular, we denote the relevant document of a query q as d^+, and the most relevant passage in d^+ as p^+. We define a counterfactual document d_i^* as the positive document d^+ with the i-th passage, i.e. p_i, changed. For the construction of the counterfactual document, we try three different ways to change the key passage: (1) **Deletion**: Removing the passage from the document; (2) **Modification**: Randomly altering some words or phrases within the passage; (3) **Replacement**: Substituting the passage with another random passage from the document.

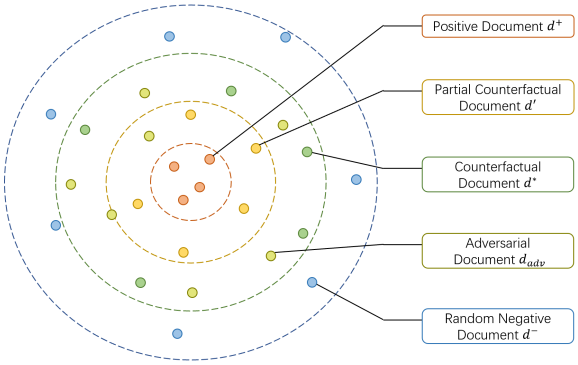

Figure 1: Illustration of the counterfactual and adversarial documents as pivots in the embedding space.

Game Theory-Based Counterfactual Learning Method. Motivated by the Shapley value in coalitional game theory (Winter, 2002), we present a novel counterfactual learning method. The extraction of key passages can be conceptualized as a coalitional game, where passages collaborate to enhance the query-document relevance. The contribution of each passage to this cooperative process reflects its importance in determining the query-document relevance. Therefore, by ranking passages based on their contributions, we can identify the most crucial passage.

Utilizing the Shapley value, we calculate the contribution of a passage using the following formula:

$$\phi_v(p_i) = \sum_{P \subset P_d \setminus \{p_i\}} \frac{|P|!(n - |P| - 1)!}{n!} (v(P \cup \{p_i\}) - v(P)), \quad (4)$$

Here, v represents the relevance score function, and $\phi_v(p_i)$ signifies the contribution of passage p_i.

This method provides interpretable measures of each passage's contribution to the overall relevance of the document. Drawing inspiration from coalitional game theory establishes a robust theoretical foundation for the counterfactual learning approach, offering a principled method for distributing gains among passages. The Shapley-based method is adaptable to various dense retrieval models, making it a versatile approach suitable for different scenarios.

3.3 Counterfactual Contrastive Learning

Based on the key passage extraction method, we aim to design a pretraining task where the model learns the relative positions of counterfactual documents on the embedding space from dense retrievers.

For a query-document pair (q, d^+), we define

its positive passage as p^+, the corresponding counterfactual document of d^+ as d^*, and a set of N negative documents as D^-.

A dense retriever learns query and document representations such that the following inequality between positive document d^+ and negative document d^- holds:

$$f(q, d^+) > f(q, d^-). \tag{5}$$

To specify the relative positions of counterfactual passage in the embedding space, one must take into account the their relevance between the positive document (i.e. source document) and the negative document.

As illustrated in Figure 1, we examine our counterfactual documents from different perspectives, establishing a cascaded learning process. Due to the substantial semantic difference between d^+ and d^*, we introduce a partial counterfactual document d' by randomly removing one sentence from p^+ in d^+. This design captures a more fine-grained semantic change. Consequently, we can get the relationship between embedding similarities:

$$f(q, d^+) > f(q, d') > f(q, d^*). \tag{6}$$

Learning to discriminate d^+, d', d^* would minimize the mutual information between representations of query and irrelevant passages in d^+, strengthening causal effects of the relevant passage p^+.

Since a counterfactual document d^* is a minimally different to the document that retains most of the semantics in d^+, d^* can be seen as a pseudo-positive example in the document retrieval. Semantic relevance between q and d^* distinguishes d^* from other negatives d^-, which provide noisy contexts with respect to q. Thus, the following holds for all d^*:

$$f(q, d^*) > f(q, d^-). \tag{7}$$

Adversarial attacks offer another means of constructing a different type of counterfactual documents. The aforementioned construction of counterfactual documents changes a continuous range of local information, while the adversarial examples changes different part of the document globally. Specifically, we construct the adversarial examples d_{adv} as:

$$d_{adv} = arg \max_{d' \in \mathbb{B}(d,\epsilon)} \left(f(q, d') - f(q, d) \right), \tag{8}$$

where $\mathbb{B}(d, \epsilon)$ signifies that a ratio of ϵ words in d are randomly replaced. Similarly, the following holds for all d_{adv}:

$$f(q, d^+) > f(q, d_{adv}) > f(q, d^-). \tag{9}$$

Building upon the aforementioned relationships, we propose a counterfactual contrastive learning method for a more robust dense retriever. Specifically, we augment the loss function of the dense retriever (Equation 3) by introducing additional counterfactual contrastive loss terms. These terms leverage counterfactual documents as pivotal elements between positives and negatives.

Classic Dense Retrieval Loss: L_{cla}, is a pairwise loss as the classic ranking loss, defined in Equation 3.

Counterfactuals as Hard Negatives Loss: L_neg, is optimized to maximize the similarity between $(q, d), (q, d')$ while minimizing the similarity between $(q, d'), (q, d^*)$. It imposes the key constraint on q, d', d^* from Equation 6 to discriminate positive documents from counterfactual documents:

$$L_{neg} = -log \frac{e^{f(q,d^+)}}{e^{f(q,d^+)} + e^{f(q,d')}} - log \frac{e^{f(q,d')}}{e^{f(q,d')} + e^{f(q,d^*)}}. \tag{10}$$

Counterfactuals as Pseudo Positives Loss: L_pos, is optimized to maximize the relative similarity between q and d^* with respect to negative documents d_j^-. The key difference between L_2 and L_3 is that the counterfactual passage is used as a positive in L_2 to retain semantic relevance in the learned embeddings.

$$L_{pos} = -log \frac{e(f(q, d^*)}{e^{f(q,d^*)} + \sum_{j=1}^{N} e^{f(q,d_j^-)}}. \tag{11}$$

Counterfactuals with Adversarial Attack Loss: L_{adv} is optimized to maximize the similarity between $(q, d^+), (q, d_{adv})$ while minimizing the similarity between $(q, d_{adv}), (q, d^*)$.

$$L_{adv} = -log \frac{e^{f(q,d^+)}}{e^{f(q,d^+)} + e^{f(q,d_{adv})}} - log \frac{e^{f(q,d_{adv})}}{e^{f(q,d_{adv})} + e^{f(q,d^*)}}. \tag{12}$$

The final loss function L is a weighted sum of all three loss fuctions:

$$L = L_{cla} + \alpha(L_{neg} + L_{adv}) + \beta L_{pos}, \tag{13}$$

where α, β are hyperparameters that determine the importance of the terms. Various strategies exist for setting these hyperparameters based on the documents and passages. In our approach, we explore three different methods:

- Relevance score (rel): Setting $\alpha = 1 - r, \beta = r, r = f(q, d^*)$.
- Shapley value (shapley): Setting $\alpha = 1 - s', \beta = s'$, where s' is the normalized shapley value.
- Couple learning (CP): Employing the couple learning method (Liu et al., 2023e) to dynamically adjust the hyperparameters.

4 Experiments

4.1 Dataset Selection and Preprocessing

In all our experiments, we utilize the MSMARCO-doc and MSMARCO-NQ dataset (Nguyen et al., 2016) for evaluating model performance. These datasets comprise extensive anonymized questions sampled from Bing's search query logs and texts extracted from 3,563,535 web pages retrieved by Bing. Relevant documents for each query are identified based on the presence of at least one related passage. By matching MSMARCO-doc dataset with MSMARCO-NQ dataset, we curate matching triples consisting of <query, document, relevant passage>, resulting in a total of 164,190 instances.

For the counterfactual contrastive learning task, we explore two training modes or setups. In the first setting, we take the aforementioned matched triples as training data. Since the passage-level relevant labels are usually hard to get, costing a lot of time and money, we try to use the LLM to assist the relevant passage labeling. In the second setting, we use gpt3.5 to label the positive passage in each <query, document> pair, simulating a scenario with a lack of labeled data. The segmentation of the documents and the annotation of the relevant passages are detailed in the appendix.

4.2 Baselines and Evaluation Metrics

Key Passage Extraction We consider two intuitive counterfactual methods for the comparison of model-agnostic key passage extraction methods:

- Document Ranking Change (δ_{rank}): We measure the change in ranking of the modified document compared to the original document in the retrieval model's results. A higher δ_{rank} signifies a more crucial passage for the query.
- Relevance Score Change (δ_{rel}): We calculate the differences in the relevance scores assigned by the retrieval model to the original and counterfactual documents. A larger δ_{rel} indicates greater passage importance.

Based on δ_{rank} and δ_{rel}, we can rerank the passages and obtain the most important passage within the document d according to the similarity scores.

Our backbone models for the counterfactual learning method include two kinds of document retrieval models: (1) sparse document retrieval models: BM25 (Robertson et al., 2009), docT5query (Nogueira et al., 2019), deepCT (Dai and Callan, 2020); (2) dense document retrieval models: DPR (Karpukhin et al., 2020), ANCE (FirstP) (Xiong et al., 2020), colBERT (Santhanam et al., 2022), ME-BERT (Luan et al., 2021).

Adversarial Attack We take four different attack methods for testing the robustness of the models: Term spamming (TS) (Gyöngyi et al., 2005) randomly replaces words with terms randomly sampled from the target query. (2) PRADA (Wu et al., 2023b) is a decision-based black-box ranking attack method against NRMs via word substitution. (3) PAT (Liu et al., 2022) is an anchor-based ranking attack method against NRMs via trigger generation. (4) MCARA (Liu et al., 2023c) formalizes attacks on DR models as a contrastive learning problem in a multi-view representation space.

The baselines of the adversarial attack models are: (1) Adversarial training (AT): We follow the vanilla AT method (Goodfellow et al., 2014) to directly include the adversarial examples during training. (2) CertDR (Wu et al., 2022) is a certified defense method for NRMs. (3) PIAT (Liu et al., 2023e) is a novel perturbation-invariant adversarial training method.

Evaluation Metrics We evaluate the model's document retrieval performance by calculating the Mean Reciprocal Rank at the top 10 positions (MRR@10d) for the retrieved documents. Simultaneously, the model's ability to extract the key passages is assessed by calculating the Mean Reciprocal Rank at the top 10 positions, which is denoted using MRR@10p in this work, for the retrieved passages within the document. In particular, this metric quantifies the model's ability to rank relevant passages highly within the top 10 results. Statistical significance is tested using the permutation test with $p < 0.05$.

4.3 Research Questions

We aim to explore the following two research questions:

1. How can counterfactual learning methods con-

Table 1: The performance of retrieving passage within the document with different conterfactual learning methods. The evaluation metric is MRR@10p. $*$ indicates significant improvements over δ_{rank} and δ_{rel} ($p < 0.05$).

methods		δ_{rank}	δ_{rel}	shapley-value		single passage	guideline
				no-overlap	merge		
sparse	BM25	0.411	0.527	0.532	**0.537***	0.374	/
	docT5query	0.442	0.568	0.555	**0.561***	0.501	/
	deepCT	0.448	0.575	0.558	**0.567***	0.497	/
dense	DPR	0.477	0.613	0.621	**0.625***	0.605	0.703
	ANCE	0.486	0.632	0.633	**0.638***	0.611	0.725
	colBERT	0.482	0.625	0.642	**0.647***	0.623	0.733
	ME-BERT	0.478	0.614	0.624	**0.628***	0.608	0.706

tribute to the prediction of crucial segments within documents?

2. How can counterfactual learning be incorporated into pretraining tasks to augment the capabilities of document retrieval models?

These questions form the core of our investigation, delving into the potential impact of counterfactual learning on the identification of significant document segments, and its broader integration into the pretraining process to improve document retrieval model capabilities.

5 Exprtimental Results

In this section, we delve into the experimental results, findings, and the performance evaluation of both the counterfactual method and the contrastive task discussed in the earlier sections of the paper.

5.1 Counterfactual Learning Methods

To address **RQ1** - *How can counterfactual learning methods contribute to the prediction of crucial segments within documents?*, we conduct experiments with various settings of counterfactual documents and different counterfactual learning methods.

Counterfactual Learning Methods We compare the performance of different counterfactual learning methods. Table 1 displays the results concerning the retrieval of positive passages within documents. The "single passage" column involves treating the target passage as a short document and re-ranking passages directly based on relevance scores from document retrieval models. The "guideline" column represents testing the passage retrieval models trained with the MSMARCO-passage dataset, serving as a performance upper bound since these models are trained with the ground truth passage relevance labels.

The δ_{rank} method exhibits suboptimal performance, attributed to coarse rank differences that sometimes result in identical rank changes for different passages. This inadequacy provides an insufficient signal to discriminate passage importance. Furthermore, rank differences are susceptible to influence from other documents in the list, introducing additional noise to the score. In contrast, the δ_{rel} method outperforms δ_{rank} by capturing relevance changes in the relevance score. Importantly, it is not influenced by other documents in the set, focusing solely on the current document.

The Shapley methods demonstrate superior performance, suggesting that Shapley values provide a better estimation of passage importance and capture more subtle differences between the original document and counterfactual document compared to δ_{rel}. This also indicates that our method of reducing computational complexity effectively estimates the Shapley value. Notably, the Shapley methods surpass the single passage approach, highlighting the value of our counterfactual learning approach.

Between the two methods of Shapley value computation, the merge method performs better than the no-overlap method. This is because the no-overlap method only takes into account half of the documents, resulting in information loss.

5.2 Counterfactual Contrastive Learning

To address **RQ2** - *How can counterfactual learning be incorporated into pretraining tasks to augment the capabilities of document retrieval models?*, we conduct experiments employing different strategies for loss weight and assess the model's robustness under various adversarial attack methods.

Retrieval Performance Table 2 depicts the performance of retrained models in key passage extraction, considering distinct loss weight strategies and various passage label settings.

Table 2: The result of the key passage extraction of the counterfactual contrastive learning task. The evaluation metric is MRR@10p. $*$ indicates significant improvements over the original model ($p < 0.05$).

	loss weight strategy	MS-MARCO passage label			gpt3.5 passage label		
		δ_{rank}	δ_{rel}	shapley (merge)	δ_{rank}	δ_{rel}	shapley (merge)
$\text{DPR}_{counter}$	relevance score	0.486	0.619	0.629	0.474	0.613	0.623
	shapley value	0.489	0.624	0.633	0.479	0.616	0.625
	coupling learning	**0.491**$*$	**0.628**$*$	**0.636**$*$	**0.483**	**0.619**	**0.629**
original DPR		0.477	0.613	0.625	0.477	0.613	0.625
$\text{ANCE}_{counter}$	relevance score	0.489	0.640	0.641	0.483	0.630	0.635
	shapley value	0.493	0.642	0.644	0.485	0.632	0.636
	coupling learning	**0.499**$*$	**0.645**$*$	**0.649**$*$	**0.489**	**0.637**	**0.640**
original ANCE		0.486	0.632	0.638	0.486	0.632	0.638

Table 3: The performance of the counterfactual contrastive learning task. The evaluation metric is MRR@10d and NDCG@10d.

	MSMARCO DOC MRR@10	TREC DL2019 DOC NDCG@10
DPR	0.277	0.554
$\text{DPR}_{counter}$	**0.280**	**0.555**
ANCE	0.315	0.622
$\text{ANCE}_{counter}$	**0.317**	**0.624**

Results from models trained with MSMARCO passage labels indicate that the coupling learning method (Liu et al., 2023e) yields the best performance. This is attributed to its dynamic adjustment of hyperparameters during the training process. The observed improvement in extracting key passages suggests that our contrastive learning methods enhance the counterfactual learning ability of dense retrieval models.

Utilizing Shapley value as the loss weight outperforms using relevance scores as the loss weight, reinforcing that Shapley value better captures the local and global differences between (d, d^*) compared to the relevance score.

Recognizing the challenge of obtaining high-quality passage-level relevance labels and the associated time and cost implications, we also conduct experiments generated by gpt-3.5 (Achiam et al., 2023). In this scenario, relevant passages in the document are labeled by gpt-3.5 using a few-shot prompt. Results demonstrate that even with this training data, our method achieves a modest performance improvement, underscoring the effectiveness of our pretraining task.

Table 3 presents the document retrieval performance of the retrained models. The results show slight improvement in document retrieval compared to the original models, though not statistically significant. This indicates that our contrastive learning task does not compromise the document retrieval ability while enhancing the capacity to extract key passages.

Model Robustness Table 4 presents the results of the retrained models' performance in extracting key passages under various adversarial attack methods. When compared with the original documents, those modified by the attack methods exhibit a decrease in model performance. Under all four attack methods, both DPR and ANCE show a lesser performance decrease after the contrastive learning task. Even under the state-of-the-art attacking method on dense retrieval models—MCARA, the retrained model demonstrates robustness. This suggests that our contrastive learning task improves the models' robustness in extracting key passages.

Table 5 presents the robustness analysis, showcasing the decline in retrieval performance under four distinct adversarial attack methods. Notably, our counterfactual learning model consistently outperforms AT and CertDR across all attack methods, and it surpasses PIAT under TS and PRADA. This observation underscores the effectiveness of our counterfactual contrastive learning model in enhancing overall model robustness. PIAT performs exceptionally well under PAT and MCARA, tailored for intricate attack methods and requiring specific attack-generated training data. In contrast, our model attains comparable performance without the need for complex adversarial examples in the training dataset.

In Table 6, we conduct an ablation study on the robustness of key passage extraction under dense retrieval attack methods. The initial row represents the attack outcomes for the original DPR

Table 4: The retrained models' performance on extracting the key passage under the dense retrieval attack methods. The evaluation metric is MRR@10p.

	Original documents	Attack method			
		TS	PRADA	PAT	MCARA
DPR	0.613	0.584 (-4.7%)	0.562 (-8.3%)	0.565 (-7.8%)	0.549 (-10.4%)
$\text{DPR}_{counter}$	0.626	0.603 (**-3.7%**)	0.587 (**-6.2%**)	0.588 (**-6.1%**)	0.576 (**-8.0%**)
ANCE	0.632	0.603 (-4.6%)	0.586 (-7.3%)	0.584 (-7.6%)	0.570 (-9.8%)
$\text{ANCE}_{counter}$	0.641	0.614 (**-4.2%**)	0.606 (**-5.5%**)	0.603 (**-5.9%**)	0.589 (**-8.1%**)

Table 5: The decrease of the retrieval performance under four adversarial attack methods. $*$ indicates significant improvements over other models ($p < 0.05$).

model	Percentage decrease of the MRR@10d			
	TS	PRADA	PAT	MCARA
AT	-13.5%	-15.8%	-16.9%	-19.2%
CertDR	-12.4%	-14.1%	-14.7%	-17.3%
PIAT	-10.7%	-12.8%	**-12.5%**$*$	**-14.8%**$*$
$\text{DPR}_{counter}$	**-10.1%**	**-12.3%**	-13.7%	-16.1%

Table 6: The ablation study of the robustness on extracting the key passage under the dense retrieval attack methods.

Training Loss	Percentage decrease of the MRR@10p			
	TS	PRADA	PAT	MCARA
L_{nat}	-4.70%	-8.30%	-7.80%	-10.40%
$+L_{pos}$	-7.10%	-9.90%	-9.30%	-12.10%
$+L_{neg}$	-5.60%	-9.00%	-8.50%	-11.10%
$+L_{pos} + L_{neg}$	-4.50%	-7.70%	-7.40%	-9.80%
$+L_{adv}$	-4.20%	-7.10%	-6.90%	-9.10%
Total	**-3.70%**	**-6.20%**	**-6.10%**	**-8.00%**

model, while subsequent rows depict results considering different components of the training loss. Results solely from $+L_{pos}$ or $+L_{neg}$ exhibit subpar performance, focusing only on a segment of the contrastive direction within the counterfactual document, thereby compromising model robustness. The combination of $L_{pos} + L_{neg}$ demonstrates marginal robustness improvement, highlighting the positive contribution of the counterfactual contrastive loss. Importantly, the inclusion of L_{adv} yields the most significant enhancement in robustness, underscoring the pivotal role of comparing the adversarial attack document with the counterfactual document. Comprehensive consideration of all loss components collectively achieves the optimum performance in extracting key passages under dense retrieval attack methods.

6 Conclusion

In conclusion, our research addresses critical challenges in dense retrieval models, focusing on the issues of key passage extraction and the vulnerability to adversarial attacks. We have emphasized the importance of enhancing model sensitivity to fine-grained relevance matching signals as a key strategy to mitigate these challenges.

By advocating for sensitivity to modifications in key passages, our proposed counterfactual regularization methods aim to render dense retrieval models more interpretable and robust. Our methodology introduces a cooperative game theory-based counterfactual passage extraction method. The demonstrated robustness against adversarial attacks underscores the potential of our approach to enhance model reliability in real-world applications.

Experimentally, our approach showcases promising results, revealing the model's capacity to learn key passages without explicit passage-level relevance annotations. More importantly, our regularized dense retrieval models exhibit superior resistance to adversarial attacks, outperforming existing anti-attack methods.

In light of these outcomes, our work provides a pathway to more interpretable and robust dense retrieval models. As we continue to grapple with the complexities of vast document corpora and evolving search demands, our research lays the groundwork for further improvements in model reliability and applicability.

7 Limitations

Scalability. The efficiency and scalability of our approach when applied to extensive document corpora warrant careful consideration. Addressing potential computational challenges for large-scale deployments is an area for further investigation.

Attack Strategies Evolution. As adversarial attack strategies evolve, there may be limitations to

the robustness of our model against novel and sophisticated attacks. Continuous monitoring and adaptation strategies may be needed.

Model Robustness Trade-offs. While our counterfactual regularization methods enhance model's robustness, there might be trade-offs with retrieval performance. Striking the right balance between robustness and performance is an ongoing challenge.

8 Ethics Statement

For an input query, our counterfactual learning method enables the direct retrieval of the key passage in the relevant document, and our counterfactual constrastive learning task improves the robustness of the dense retrieval models. However, the performance of our model is still not perfect, and it may retrieve incorrect triplets in response to given queries. Therefore, for the high-risk domains, such as biomedicine, our model should be carefully used, and it might be required to analyze retrieved docuemnts and passages before making the critical decision.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Xuanang Chen, Ben He, Zheng Ye, Le Sun, and Yingfei Sun. 2023. Towards imperceptible document manipulations against neural ranking models. *arXiv preprint arXiv:2305.01860*.

Seungtaek Choi, Haeju Park, Jinyoung Yeo, and Seungwon Hwang. 2020. Less is more: Attention supervision with counterfactuals for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6695–6704. Association for Computational Linguistics.

Zhuyun Dai and Jamie Callan. 2020. Context-aware term weighting for first stage passage retrieval. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 1533–1536.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual visual explanations. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2376–2384. PMLR.

Zoltán Gyöngyi, Hector Garcia-Molina, et al. 2005. Web spam taxonomy. In *AIRWeb*, volume 5, pages 39–47.

Fredrik D. Johansson, Uri Shalit, and David A. Sontag. 2016. Learning representations for counterfactual inference. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 3020–3029. JMLR.org.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 39–48. ACM.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4066–4076.

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020. Pre-training via paraphrasing. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ke Liang, Yue Liu, Sihang Zhou, Wenxuan Tu, Yi Wen, Xihong Yang, Xiangjun Dong, and Xinwang Liu. 2024. Knowledge graph contrastive learning based on relation-symmetrical structure. *IEEE Trans. Knowl. Data Eng.*, 36(1):226–238.

Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, and Xinwang Liu. 2023. Learn from relational correlations and periodic

events for temporal knowledge graph reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 1559–1568. ACM.

Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. 2020. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3285–3292. Association for Computational Linguistics.

Jiawei Liu, Yangyang Kang, Di Tang, Kaisong Song, Changlong Sun, Xiaofeng Wang, Wei Lu, and Xiaozhong Liu. 2022. Order-disorder: Imitation adversarial attacks for black-box neural ranking models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2025–2039.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Wei Chen, and Xueqi Cheng. 2023a. On the robustness of generative retrieval models: An out-of-distribution perspective. *CoRR*, abs/2306.12756.

Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng. 2023b. Black-box adversarial attacks against dense retrieval models: A multi-view contrastive learning method. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, pages 1647–1656. ACM.

Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng. 2023c. Black-box adversarial attacks against dense retrieval models: A multi-view contrastive learning method. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1647–1656.

Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng. 2023d. Topic-oriented adversarial attacks against black-box neural ranking models. *arXiv preprint arXiv:2304.14867*.

Yu-An Liu, Ruqing Zhang, Mingkun Zhang, Wei Chen, Maarten de Rijke, Jiafeng Guo, and Xueqi Cheng. 2023e. Perturbation-invariant adversarial training for neural ranking models: Improving the effectiveness-robustness trade-off. *arXiv preprint arXiv:2312.10329*.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual VQA: A cause-effect look at language bias. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12700–12710. Computer Vision Foundation / IEEE.

Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019. From doc2query to doctttttquery. *Online preprint*, 6:2.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. 2018. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. New Jersey, USA.

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

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

Yongho Song, Dahyun Lee, Kyungjae Lee, and Jinyeong Yeo. 2023. Revisiting dense retrieval with unanswerable counterfactuals. *arXiv preprint arXiv:2304.03031*.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Bing Tian, Yixin Cao, Yong Zhang, and Chunxiao Xing. 2022. Debiasing NLU models via causal intervention and counterfactual reasoning. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11376–11384. AAAI Press.

Shaokun Wang, Tian Gan, Yuan Liu, Jianlong Wu, Yuan Cheng, and Liqiang Nie. 2023. Micro-influencer recommendation by multi-perspective account representation learning. *IEEE Trans. Multim.*, 25:2749–2760.

Shaokun Wang, Tian Gan, Yuan Liu, Li Zhang, Jianlong Wu, and Liqiang Nie. 2022. Discover micro-influencers for brands via better understanding. *IEEE Trans. Multim.*, 24:2595–2605.

Eyal Winter. 2002. The shapley value. *Handbook of game theory with economic applications*, 3:2025–2054.

Chen Wu, Ruqing Zhang, Jiafeng Guo, Wei Chen, Yixing Fan, Maarten de Rijke, and Xueqi Cheng. 2022. Certified robustness to word substitution ranking attack for neural ranking models. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2128–2137.

Chen Wu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2023a. PRADA: practical black-box adversarial attacks against neural ranking models. *ACM Trans. Inf. Syst.*, 41(4):89:1–89:27.

Chen Wu, Ruqing Zhang, Jiafeng Guo, Maarten De Rijke, Yixing Fan, and Xueqi Cheng. 2023b. Prada: practical black-box adversarial attacks against neural ranking models. *ACM Transactions on Information Systems*, 41(4):1–27.

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*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021a. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Wenhan Xiong, Xiang Lorraine Li, Srini Iyer, Jingfei Du, Patrick S. H. Lewis, William Yang Wang, Yashar Mehdad, Scott Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oguz. 2021b. Answering complex open-domain questions with multi-hop dense retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Table 7: The retrieval effectiveness of retrieving passage within the document according to δ_{rel} with difference window size and different type of counterfactual construction methods. The evaluation metric is MRR@10p. The best performance among various counterfactual document construction methods for a model is boldfaced.

	Models	window size=128			window size=64		
		deletion	modification	replacement	deletion	modification	replacement
sparse	BM25	**0.527**	0.501	0.513	0.491	0.476	0.485
	docT5query	**0.568**	0.543	0.552	0.524	0.505	0.517
	deepCT	**0.575**	0.550	0.559	0.538	0.511	0.523
dense	DPR	**0.613**	0.597	0.604	0.589	0.565	0.576
	ANCE	**0.632**	0.609	0.620	0.606	0.587	0.594
	colBERT	**0.625**	0.604	0.611	0.601	0.581	0.589
	ME-BERT	**0.614**	0.595	0.601	0.591	0.568	0.577

Table 8: The statistics of the segmented documents.

window size	positive psg ratio	average psg num
64	51.80%	42.6
128	95.80%	21.6

9 Appendix

Document Segmentation To facilitate further analysis, we segment the documents in the dataset using a fixed-size sliding window approach. Two window sizes, 64 and 128, were utilized. This segmentation was performed with a 50% overlap. From the segmented document sections, we regard the segment as positive with the highest overlap with the relevant passage in each triple. This segment is then annotated as the positive example which can be used for both training and evaluation processes. The statistics of the segmented documents are presented in Table 8. The "positive passage ratio" denotes the ratio of the original labeled passage fully covered in one segment after segmentation. With a window size of 128, 95.8% of documents maintain the labeled passage intact after segmentation. The "average passage num" signifies the average number of passages in each document.

Since the definition of the Shapley value requires the passages to be non-overlapping we take two different methods to get the final Shapley value:

- **Non-overlap**: In this approach, only half of the segmented passages, which do not overlap with each other, are taken into account.
- **Merge**: In this approach, the Shapley value of two groups of non-overlapping passages is independently calculated, and the final Shapley value is obtained as the moving average of three consecutive passages.

Counterfactual Document Construction We explore the optimal settings for constructing the counterfactual documents. Table 7 illustrates the results obtained with three types of counterfactual document construction methods and each is experimented with two different segmentatin window sizes. The δ_{rel} method results are showcased, with δ_{rank} and shapley exhibiting similar outcomes.

From Table7, we observe that the performance of the models with the window size of 64 performs poorly due to only 51.8% of the positive passages remaining intact, leading to significant information loss and a reduction in the counterfactual document's quality. In contrast, a window size of 128 preserves 95.8% of positive passages, providing a conducive environment for the model to extract significant passages effectively. Among the three construction types, the deletion method consistently yields the best performance because deletion induces the most substantial information change. Conversely, the modification method, involving minor word substitutions, results in counterfactual documents too similar to the originals, thus showing the poorest performance.

Consequently, we set the window size as 128 and the construction type as deletion for subsequent experiments.