On the Scaling of Robustness and Effectiveness in Dense Retrieval

Yu-An Liu
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China
liuyuan21b@ict.ac.cn

Ruqing Zhang*
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China
zhangruqing@ict.ac.cn

Jiafeng Guo*
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China
guojiafeng@ict.ac.cn

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
m.derijke@uva.nl

Yixing Fan
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China
fanyixing@ict.ac.cn

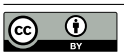
Xueqi Cheng
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China
cxq@ict.ac.cn

Abstract

Robustness and Effectiveness are critical aspects of developing dense retrieval models for real-world applications. It is known that there is a trade-off between the two. Recent work has addressed scaling laws of effectiveness in dense retrieval, revealing a power-law relationship between effectiveness and the size of models and data. Does robustness follow scaling laws too? If so, can scaling improve both robustness and effectiveness together, or do they remain locked in a trade-off?

To answer these questions, we conduct a comprehensive experimental study. We find that: (i) Robustness, including out-of-distribution and adversarial robustness, also follows a scaling law. (ii) Robustness and effectiveness exhibit different scaling patterns, leading to significant resource costs when jointly improving both. Given these findings, we shift to the third factor that affects model performance, namely the optimization strategy, beyond the model size and data size. We find that: (i) By fitting different optimization strategies, the joint performance of robustness and effectiveness traces out a Pareto frontier. (ii) When the optimization strategy strays from Pareto efficiency, the joint performance scales in a sub-optimal direction. (iii) By adjusting the optimization weights to fit the Pareto efficiency, we can achieve Pareto training, where the scaling of joint performance becomes most efficient. Even without requiring additional resources, Pareto training is comparable to the performance of scaling resources several times under optimization strategies that overly prioritize either robustness or effectiveness. Finally, we demonstrate that our findings can help deploy dense retrieval models in real-world applications that scale efficiently and are balanced for robustness and effectiveness.

*Jiafeng Guo and Ruqing Zhang are the corresponding authors.



This work is licensed under a Creative Commons Attribution 4.0 International License.

SIGIR '25, Padua, Italy

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730049

CCS Concepts

• **Information systems** → **Retrieval models and ranking**.

Keywords

Dense retrieval, Robustness, Effectiveness, Neural scaling law

ACM Reference Format:
Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2025. On the Scaling of Robustness and Effectiveness in Dense Retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3726302.3730049

1 Introduction

Dense retrieval models have achieved state-of-the-art effectiveness, which reflects overall performance under *normal conditions*. But they inherit vulnerabilities commonly associated with general neural networks [24, 37, 55], making them inherently disadvantaged in terms of robustness in *abnormal situations*, such as handling out-of-distribution (OOD) data or adversarial attacks [26, 35, 41, 46]. Due to this limitation, dense retrieval models face a trade-off between robustness and effectiveness [35, 41, 54, 55]. Understanding the interplay between these two aspects is critical for ensuring reliable ranking performance across diverse practical scenarios.

We explore robustness and effectiveness from the perspective of scaling, aiming to guide the development of dense retrieval models that excel in both dimensions. Recent studies [6, 32] have examined scaling laws of effectiveness in dense retrieval, revealing that scaling model size and data volume enhance effectiveness and optimize the training process. These findings motivate us to investigate whether scaling can simultaneously improve both robustness and effectiveness: *Does robustness follow similar scaling laws?*

Scaling laws of robustness. We address this question through experimental analysis, starting with an investigation of *out-of-distribution (OOD) robustness* [41] in dense retrieval. Our analysis focuses on the impact of two key factors, i.e., *model size* and *dataset*

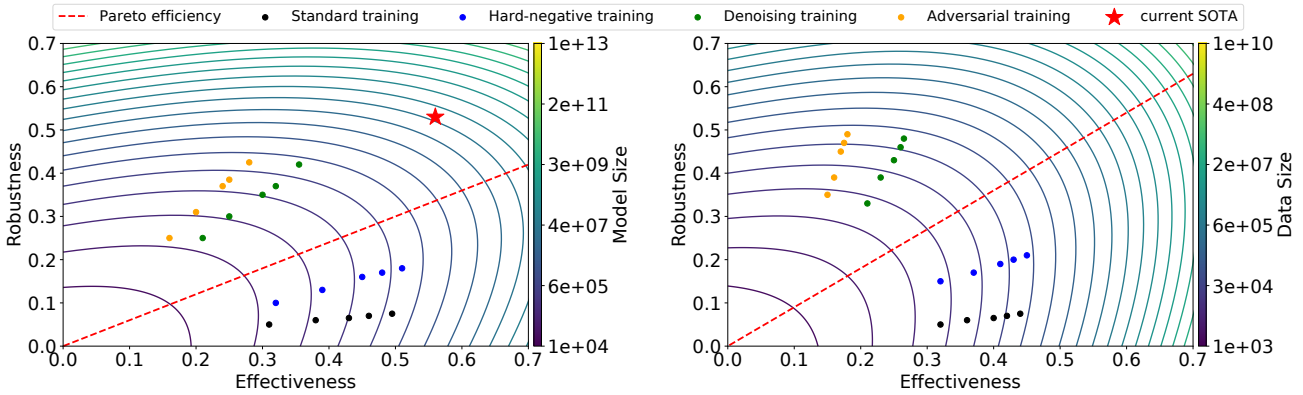


Figure 1: Joint scaling trends in robustness and effectiveness of the BERT model with respect to model size and data size across different optimization strategies. Effectiveness is evaluated on MS MARCO, and robustness is the average of the OOD robustness on BEIR and the adversarial robustness against ranking attacks on MS MARCO. The metric in the figure is obtained by inverse normalization of the (average) contrastive entropy. More experimental details are in Section 4.

size, on robustness performance. Following [6], we employ the contrastive entropy metric to assess the performance of dense retrieval models. Experiments are conducted using dense retrieval models implemented with various pre-trained language models, with non-embedding parameter sizes ranging from 0.5 to 87 million. These models are trained on Chinese and English web search datasets, with training dataset sizes varying from 30K to 480K. Our results show that robustness also adheres to a precise power-law scaling relationship w.r.t. both model size and dataset size. We also observe that the scaling law for dataset size remain consistent across different annotation qualities. We investigate another critical aspect of robustness in IR scenarios [2, 19, 46]: *adversarial robustness*. While scaling laws show variability (or "noise") in adversarial robustness for the original model, they are precise for dense retrieval models after adversarial training [27].

Different scaling patterns between robustness and effectiveness. With the scaling laws that govern robustness and effectiveness in dense retrieval, we aim to explore: *Whether robustness and effectiveness can be improved together through scaling*. The answer is: *not exactly*. The robustness of dense retrieval models is more sensitive to dataset size, while their effectiveness is more influenced by model size. This re-introduces a trade-off between robustness and effectiveness in terms of resource requirements, as jointly improving requires substantial parameters and data costs. E.g., our experiments show that achieving a 10% improvement in both robustness and effectiveness for the state-of-the-art dense retrieval model (Llama2Vec, 7B) [21] requires both a model size comparable to GPT-4 (175B) [34] and scaling up training data 10-fold.

Building on these findings, we consider a third critical factor influencing model performance, namely the optimization strategy, beyond just data and model. By evaluating the performance of optimization strategies, our goal is to identify approaches that can jointly improve robustness and effectiveness, even within constrained resource budgets, rather than relying solely on scaling up. Our analysis reveals three key findings:

Pareto efficiency exists in robustness and effectiveness. We keep the model and dataset size constant while exploring techniques such as hard-negative training [50], denoising training [3], and adversarial training [27] to evaluate their impact on both robustness and effectiveness. As shown in Figure 1, fitting robustness and effectiveness under various optimization strategies reveals a Pareto frontier for joint performance. This represents *Pareto efficiency* [44], the red dashed line, where robustness cannot be better off without making effectiveness worse off, and vice versa.

Deviation from Pareto efficiency leads to suboptimal scaling. As shown in Figure 1, most representative training methods deviate from Pareto-efficient training by overemphasizing either robustness or effectiveness. Robustness and effectiveness are not appropriately assigned in these optimization objectives, leading to a sub-optimal scaling direction with inefficient performance improvement. Take hard-negative training as an example: its overemphasis on model effectiveness on hard negatives limits the performance of robustness, resulting in attenuated gains in scaling model and data size.

Pareto training achieves balance and efficient scaling. To achieve Pareto efficiency between robustness and effectiveness, we propose a *Pareto training* method.[1] Pareto training estimates the distance between the current model state and the Pareto efficiency at each training step, and adaptively adjusts the weights of robustness and effectiveness in the training objective. Compared with standard training, Pareto training can improve scaling efficiency by up to 2.5x within a certain range. Even without scaling, its joint performance is comparable to the gains achieved by simply scaling up, such as a multi-fold increase in dataset volume or in model size.

From a joint perspective of robustness and effectiveness, we show how Pareto efficiency can be used in practice to allocate resource budgets and highlighted the resource-friendliness of Pareto training. We hope our findings provide valuable insights for developing dense retrieval models that are not only effective but also robust.

2 Problem Statement

2.1 Task Description

Dense retrieval. Dense retrieval serves as an implementation of the first-stage retrieval. Given a corpus and a query, the goal of the first-stage retrieval is to return a ranked list of top-K most relevant

[1] Our code is open-sourced at https://github.com/Davion-Liu/Robust-Effect_Scale.

documents based on the relevance score of each document in the corpus to the query [10, 16, 48, 54]. In dense retrieval, documents and queries are encoded into dense vectors; relevance is computed using these representations [54]. Given a query q and document d, dense retrieval models compute the relevance score $\mathrm{Rel}(q, d)$ as

$$\mathrm{Rel}(q, d) = f\left(\psi(q), \phi(d)\right), \tag{1}$$

where $f(\cdot)$ is a interaction function typically realized by dot product, $\psi(\cdot)$ and $\phi(\cdot)$ are functions mapping queries and documents into l-dimensional vectors, respectively [54]. In this paper, we focus on the use of shared encoders for queries and documents, a widely adopted and effective approach in dense retrieval [5, 16, 18].

Effectiveness of dense retrieval. Effectiveness refers to the average ranking performance under conditions consistent with the training data. Formally, given a dense retrieval model $f_{\mathcal{D}_{\text{train}}}$ trained on the original training set $\mathcal{D}_{\text{train}}$, its effectiveness E refer to the ranking performance $\mathcal{R}_M$ under the original test data $\mathcal{D}_{\text{test}}$:

$$E = \mathcal{R}_M\left(f_{\mathcal{D}_{\text{train}}}; \mathcal{D}_{\text{test}}\right). \tag{2}$$

Robustness of dense retrieval. In IR, robustness refers to the ability of a model to maintain ranking performance when facing unseen data from abnormal conditions. In general, abnormal conditions include OOD queries and documents [41], adversarial attacks [45], and non-retrievable documents [1]. Given a well-trained dense retrieval model $f_{\mathcal{D}_{\text{train}}}$, its robustness is derived from its ranking performance under unseen test data $\mathcal{D}^*_{\text{test}}$:

$$R = \mathcal{R}_M\left(f_{\mathcal{D}_{\text{train}}}; \mathcal{D}^*_{\text{test}}\right). \tag{3}$$

We focus on OOD robustness and adversarial robustness. We leave the exploration of other types of robustness for future work.

2.2 Training Setting

Model architecture. With the development of large-scale pre-trained language models, dense retrieval models have advanced significantly. These models mostly adopt transformer-based architectures [43]. We use the BERT [17] and ERNIE [40] families of models as our base architecture. For English benchmarks, we use 24 BERT checkpoints from Google's original release, ranging from BERT-Tiny (0.5 million parameters) to BERT-Base (82 million parameters). For Chinese benchmarks, we adopt the ERNIE series, which shares similar pre-training tasks with BERT and is trained on Chinese corpora. To investigate the impact of model size on robustness, we experiment with BERTs of various sizes.

Following [6], we initialize these models' pre-trained language models followed by fine-tuning them on annotated datasets. The output vector is typically extracted either from the [CLS] token representation or via mean pooling over the final transformer layer's outputs. To ensure comparability, a projection layer is added to each model to standardize embedding dimensions to 768.

Training data. Following [6], we use two large-scale datasets to train our dense retrieval models. (i) MS MARCO Passage Ranking [31] is an English web search dataset with about 8.8 million passages from web pages and 0.5 million training queries. (ii) T2Ranking [49] is a Chinese web search dataset with about 300k queries and over 2 million passages collected from real-world search engines. We take each query-positive-passage pair in the training data as an independent data point. To investigate the impact of data size on robustness, we experiment with various numbers of training data points. We randomly sample training data for each dataset from 30K to 480K, resulting in five sets of training data.

Evaluation data. For effectiveness, we use the dataset on which the model was trained to evaluate it. For OOD robustness, we use two benchmarks to measure the OOD robustness of the English and Chinese dense retrieval models, respectively. For adversarial robustness, the test dataset consists of adversarial samples (highly ranked documents) generated by attacking the dense retrieval model.

For OOD robustness, we adopt benchmarks spanning diverse domains, with both broad and specialized topics, varying text types, sizes, query lengths, and document lengths. The diversity of data corresponds to the OOD robustness challenges of reality. (i) For English, we adopt the BEIR benchmark [41], which includes 18 retrieval datasets with 9 different retrieval tasks. (ii) For Chinese, there is no benchmark specific to OOD robustness, so we integrated 5 retrieval datasets from different domains to construct one, named BCIR. They are e-commerce, entertainment video, and medical datasets in Multi-CPR [28], a medical community QA dataset, cMedQA2 [53], and a news retrieval dataset, TianGong-PDR [47].

Training method. For the training method, we follow [6] to adopt the most straightforward random negative sampling and in-batch negative techniques, as the *standard training* method in Section 3. Given a query-passage pair (q_i, d_i^+), we optimize the dense retrieval model with contrastive ranking loss:

$$\mathcal{L}(\theta) = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(\mathrm{Rel}\left(q_i, d_i^+; \theta\right)\right)}{\exp\left(\mathrm{Rel}\left(q_i, d_i^+; \theta\right)\right) + \sum_j \exp\left(\mathrm{Rel}\left(q_i, d_j^-; \theta\right)\right)}, \tag{4}$$

where B is the training batch size, $\{d_j^-\}$ is the set of negative passages, and $\mathrm{Rel}\left(q, d; \theta\right)$ denotes the relevance score evaluated by model f with parameters θ. We fine-tune the models for a fixed 10,000 steps and randomly sample 256 negatives at each step.

2.3 Evaluation Protocol

Evaluation for effectiveness. To evaluate the effectiveness of dense retrieval models, we follow [6] to employ contrastive entropy as our evaluation metric M, which is easy to observe trends and has been shown to perform consistently with discrete ranking metrics like NDCG@K or MRR@K. For each query-passage pair in the test set, the *contrastive entropy* is calculated as:

$$\mathrm{CE}(q_i, d_i^+; \theta) = -\log\frac{\exp\left(\mathrm{Rel}\left(q_i, d_i^+; \theta\right)\right)}{\exp\left(\mathrm{Rel}\left(q_i, d_i^+; \theta\right)\right) + \sum_j \exp\left(\mathrm{Rel}\left(q_i, d_j^-; \theta\right)\right)}, \tag{5}$$

where $\{d_j^-\}$ is the randomly selected negative set with 256 passages.

Evaluation for robustness. To simulate unseen data, we adopt zero-shot evaluation on the robustness benchmarks, using the average contrastive entropy as a measure. Given a benchmark with n test datasets, the *average contrastive entropy* across the benchmark is calculated as:

$$\mathrm{ACE}(\theta) = \frac{\sum_{i=1}^{n}\mathrm{CE}(\mathcal{D}_i; \theta)}{n}, \tag{6}$$

where $\mathcal{D}_i$ is a test dataset in the benchmark.

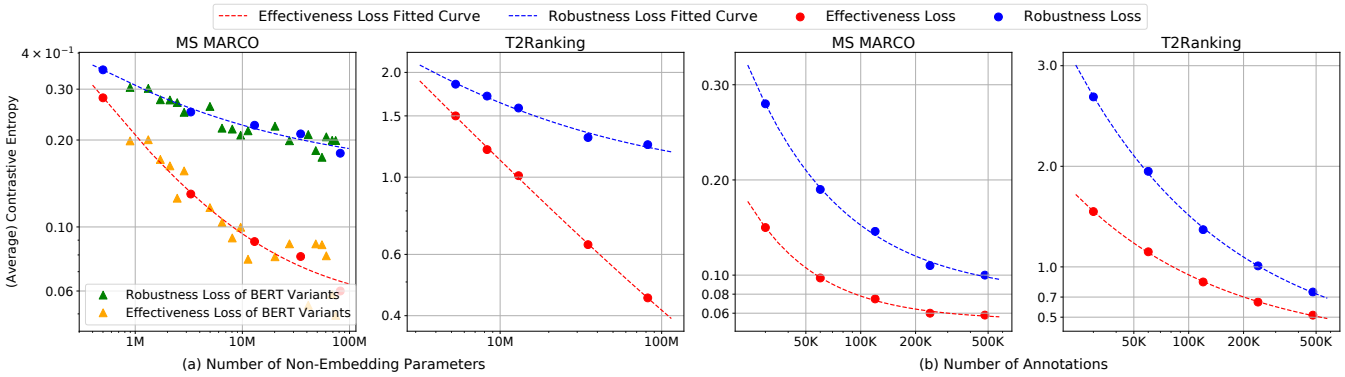


Figure 2: Scaling law for OOD robustness and effectiveness with model size (a) and data size (b) on MS MARCO and T2Ranking, respectively. Points represent the actual performance. Note that a decrease in loss represents an increase in performance.

Table 1: Fitting parameters for model size scaling.

Training dataset	Aspect	M	μ	δ_f	R^2
MS MARCO	OOD	3.70×10^4	0.55	0.05	0.997
	Effect.	3.48×10^4	0.55	0.05	0.998
T2Ranking	Robust.	4.23×10^6	0.46	0.96	0.989
	Effect.	1.12×10^7	0.48	0.07	0.999

Table 2: Fitting parameters for data size scaling.

Training dataset	Aspect	D	η	$\delta_{\mathcal{D}_{\text{train}}}$	R^2
MS MARCO	OOD	4.34×10^3	0.83	0.08	0.998
	Effect.	3.71×10^3	1.11	0.05	0.998
T2Ranking	Robust.	1.93×10^5	0.51	0.12	0.996
	Effect.	5.41×10^4	0.52	0.19	0.998

3 Scaling Laws of Robustness

In this section, we investigate the scaling laws of OOD robustness with respect to model size and data size, examine the impact of annotation quality on these scaling laws, and validate the identified scaling laws in the context of adversarial robustness.

3.1 OOD Robustness w.r.t. Model Size

Experimental setup. We fine-tune models of different sizes using the complete training set. In line with [6], to prevent underfitting or overfitting, we avoid early stopping and instead report the best results obtained on the training dataset.

Scaling law w.r.t. model size. The (average) contrastive entropy variation with model size is shown in Figure 2 (a). We see that the OOD robustness increases with the size of the model parameters. On the MS MACRO dataset, square points represent the official checkpoints of 19 differently sized BERT models. Based on our observations, we fit the scaling law of OOD robustness in terms of model sizes with log-linear functions following [6, 15]:

$$L(f) = \left(\frac{M}{|f|}\right)^{\mu} + \delta_f, \tag{7}$$

where $|f|$ represents the number of non-embedding parameters of the model, and $L(f)$ denotes the model's contrastive entropy on the test set. The parameters M, μ, and δ_f are the coefficients.

We employ the least squares method to fit the linear curve and obtain the parameters in Eq. 7 shown in Table 1. The coefficient of determination (R^2) suggests the fitting error is acceptable.

OOD robustness scales smoothly with model size. The results indicate that OOD robustness follows a precise power-law relationship with model size. The scaling behavior of OOD robustness with model size remains consistent across our Chinese and English benchmarks. When comparing the scaling laws of OOD robustness and effectiveness, we observe that effectiveness exhibits more dramatic variations with the number of model parameters, while OOD robustness changes more gradually. This suggests that larger models have a higher upper bound for ranking effectiveness. However, for OOD robustness, increasing the number of parameters may introduce more vulnerable neurons, potentially undermining the model's overall OOD robustness.

The coefficients M, μ, and δ_f are derived from the dataset, where δ_f represents the inherent loss that cannot be optimized, possibly due to incorrect or incomplete annotations. This observation also explains the higher δ_f values in the OOD robustness law, which can be attributed to the inconsistent annotation quality across datasets used in the OOD robustness benchmarks.

3.2 OOD Robustness w.r.t. Data Size

Experimental setup. We fix the model size and use different sizes of training data to construct the training set. Following [6], to avoid destabilizing effects of small-size models, we use model experiments at the largest scales in this experiment, i.e., BERT-Base and ERNIE-Base.

Scaling law w.r.t. data size. The (average) contrastive entropy variation with data size is displayed in Figure 2 (b). OOD robustness improves as the data size increases. Based on the observation, we propose to fit the scaling law of OOD robustness in terms of data size as follows:

$$L(\mathcal{D}_{\text{train}}) = \left(\frac{D}{|\mathcal{D}_{\text{train}}|}\right)^{\eta} + \delta_{\mathcal{D}_{\text{train}}}, \tag{8}$$

where $|\mathcal{D}_{\text{train}}|$ represents the number of annotated query-passage pairs, and $L(\mathcal{D}_{\text{train}})$ denotes the model's contrastive entropy on the test set. Parameters D, η, and $\delta_{\mathcal{D}_{\text{train}}}$ are the coefficients.

We employ the least squares method to fit the linear curve and obtain the parameters in Eq. 8 shown in Table 2. The coefficient of determination (R^2) suggests the fitting error is acceptable.

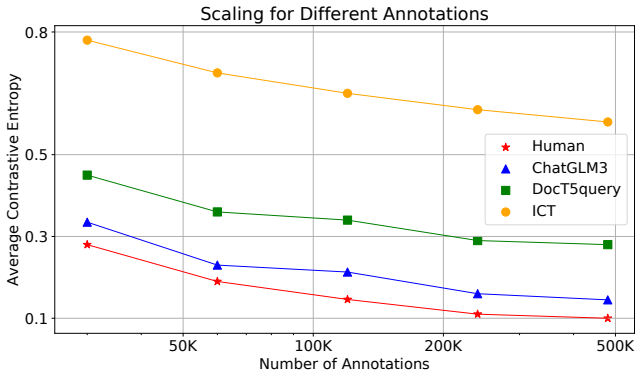


Figure 3: Scaling laws for OOD robustness with data size under different annotation methods.

OOD robustness scales dramatically with data size. The results indicate that OOD robustness follows a power-law scaling relationship with data size. The OOD robustness performance with increasing data size shows similar trends across our Chinese and English benchmarks. When comparing the scaling laws of OOD robustness and effectiveness, we observe that increasing the data size contributes more significantly to improving OOD robustness than effectiveness. This is likely because larger datasets provide a richer variety of samples, enabling models to better optimize their empirical decision boundaries. Establishing stable decision boundaries by exposing the model to sufficiently diverse data plays a crucial role in enhancing OOD robustness.

The coefficients D, η, and $\delta_{\mathcal{D}\text{train}}$ are derived from the dataset, where $\delta_{\mathcal{D}\text{train}}$ represents the inherent loss associated with the dataset. By comparing δ_f and $\delta_{\mathcal{D}\text{train}}$, we find that the inherent losses from both the model and data perspectives are nearly identical. This observation further underscores the precision of the OOD robustness scaling law we have developed.

3.3 OOD Robustness for Annotation Quality

Furthermore, we investigate whether the scaling effect for data size (Eq. 8) remains consistent across datasets of varying quality.

Experimental setup. To investigate the impact of different annotation qualities on robustness, we employ query generation techniques [33] to create three distinct types of annotations. (i) Inverse Cloze Task (ICT) [20] extracts key sentences from passages as the pseudo-query for the passage. (ii) DocT5query [33] uses a supervised generation model trained on human annotations to produce multiple queries for each passage. (iii) ChatGLM3 [51] is a large language model (LLM) which generates relevant queries for given passages. Following [6], for ICT and ChatGLM3, we generate a query for each positive document annotated by humans and for docT5query, we randomly sample 500,000 passages from the corpus for query generation, as the training data. We take the OOD robustness performance in the English benchmark as an example.

Scaling law w.r.t. data size holds across different annotation qualities. The results, shown in Figure 3, focus on the robustness of the English benchmark, with similar observations in the Chinese dataset. When manually labeled data is replaced with data annotated using different methods, the pattern of OOD robustness scaling remains consistent. Models trained with manually labeled data achieve the highest OOD robustness, suggesting that manual annotation is still the most effective approach. This finding aligns with the observations of [6] on effectiveness.

Interestingly, data annotated by generative models shows OOD robustness performance comparable to manual annotation, indicating that this method could be a cost-effective alternative for achieving satisfactory OOD robustness. This highlights the potential for exploring hybrid annotation strategies, which may reduce annotation costs while maintaining or even enhancing the OOD robustness of the trained models.

3.4 Extension to Adversarial Robustness

To validate the reliability of the scaling laws for OOD robustness with respect to model size (Eq. 7) and data size (Eq. 8), we test them in an adversarial robustness scenario.

Experimental setup. We use a representative method for attacking IR models, the word substitution ranking attack (WSRA). WSRA promotes a target document in rankings by replacing important words with synonyms [45]. We randomly sample 1,000 test queries in BEIR, for each sampled query, we randomly sample 1 document from 9 ranges in the candidates following [27, 45], i.e., $[100, 200], \ldots, [900, 1000]$, respectively. We attack these 9 target documents to achieve their corresponding adversarial examples using WSRA. Finally, we evaluate the average contrastive entropy of dense retrieval models under the attacked list with 9 adversarial examples and its query as the adversarial robustness.

In addition, we explore whether the model after adversarial training applies the scaling laws and robustness. We use the adversarial training method proposed by [27] to train our model and observe the trend of their adversarial robustness.

Adversarial robustness of the vanilla model is poorly suited for fitting. The results of scaling laws for adversarial robustness w.r.t. data and model sizes for vanilla dense retrieval models are shown in Figure 4. We take the robustness performance in the English benchmark as an example, with similar observations for the Chinese dataset. Clearly, adversarial robustness poses greater challenges for dense retrieval models than OOD robustness. While the general trend of adversarial robustness mirrors that of OOD robustness, it exhibits significant fluctuations and noise. This can be attributed to the fact that adversarial samples are carefully crafted to exploit the model's vulnerabilities, making them particularly difficult to handle. As a result, models often struggle with adversarial attacks, leading to relatively inconsistent performance. When models have not been exposed to adversarial attacks, their adversarial robustness cannot be accurately captured by a scaling law.

Adversarial robustness of the model trained with adversarial examples is easily fitted. When we examine the adversarial robustness of the dense retrieval model after adversarial training, as shown in Figure 4, adversarial training significantly enhances robustness across all model and data sizes. Surprisingly, the robustness scaling for the adversarially-trained model becomes smooth and follows clear patterns. The effectiveness and adversarial robustness of the adversarially-trained model fit the scaling law well, as shown in Table 3. This may be because the carefully crafted adversarial samples are entirely unfamiliar to the model, and after

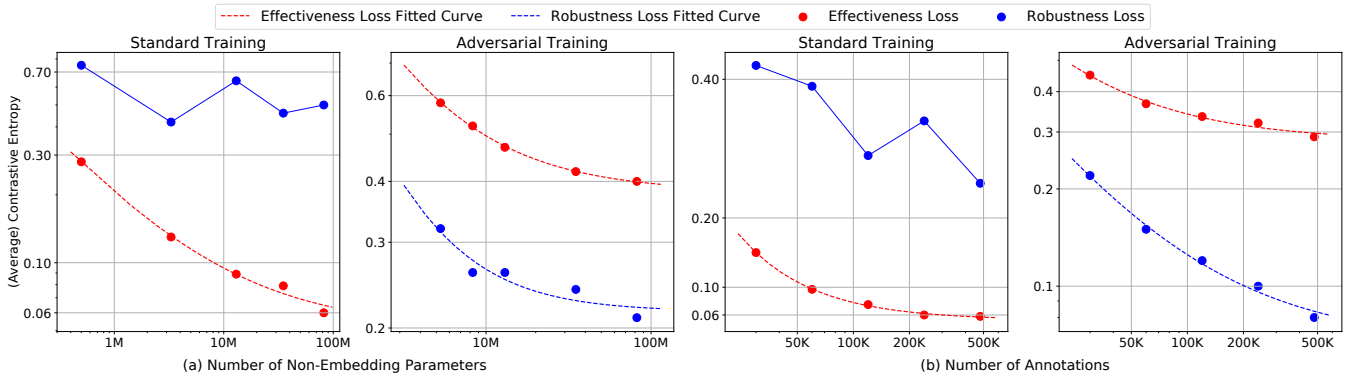


Figure 4: Scaling law for adversarial robustness and effectiveness with model size (a) and data size (b) on MS MARCO, respectively. We focus on adversarial training and standard training. The solid line indicates that it cannot be fitted into a scaling law.

Table 3: Fitting parameters for model and data size scaling on MS MARCO with adversarial robustness and effectiveness.

	Aspect	M	μ	δ_f	R^2
Model size	Adv.	6.94×10^5	1.14	0.22	0.908
	Effect.	8.34×10^5	0.87	0.38	0.999
	Aspect	D	η	$\delta_{\mathcal{D}_{\text{train}}}$	R^2
Data size	Adv.	2.81×10^3	0.80	0.07	0.995
	Effect.	3.74×10^3	0.87	0.28	0.986

encountering similar examples, the model learns to behave more consistently, revealing the underlying scaling laws. This finding aligns with the view that adversarial robustness is a form of extreme OOD robustness [9, 57]. The results suggest a general scaling law for robustness, paving the way for extending our findings to a broader and more diverse range of robustness scenarios.

3.5 Resource Budget Concerns

From our results, it is clear that there are similar scaling laws for both robustness and effectiveness. The key difference lies in their sensitivity to changes: robustness is more affected by variations in data, while effectiveness shows greater responsiveness to model size. This suggests that improving both robustness and effectiveness simultaneously comes with a significant resource overhead. Achieving substantial gains in both aspects would require a large investment in both model parameters and training data. E.g., based on our scaling laws, further improving both effectiveness and robustness by 10% from the current state-of-the-art model (Llama2Vec, 7B) [21] would necessitate a model size on par with GPT-4 (175B) [34] and scaling the training data by a factor of ten. Both demands seem prohibitively expensive and practically challenging to meet.

So far, the scaling performance of robustness and effectiveness appears to present a trade-off: within limited resource budgets, it seems feasible to make significant gains in only one aspect. This motivates us to investigate efficient approaches that could help improve the trade-off, potentially even optimizing both robustness and effectiveness jointly, rather than solely relying on scaling up.

4 Joint Scaling of Robustness and Effectiveness

In this section, we conduct an in-depth exploration of the scaling of robustness and effectiveness through their joint performance. We introduce Pareto efficiency between robustness and effectiveness, and propose a Pareto training method that uses Pareto efficiency to jointly improve robustness and effectiveness. We illustrate our findings using the performance of the BERT series on the English benchmark, with similar observations for the Chinese benchmark.

4.1 Pareto Efficiency

The performance of dense retrieval is determined by three key factors: the model, the data, and the optimization strategy. To efficiently improve the joint performance in terms of robustness and effectiveness, we explore optimization strategies and examine their impact across different models and data sizes.

Experimental setup. Besides standard training, which uses random negative sampling and in-batch negative techniques, we adopt three optimization strategies that consider robustness in training dense retrieval models: (i) hard-negative training [50] uses BM25 [38] to mine hard negative samples, which are mixed with random negative samples during training; (ii) denoising training [3] estimates noise-invariant relevance by simulating noisy documents in the training process; and (iii) adversarial training [27] enhances model robustness by modeling the boundary error between natural ranking and the ranking under adversarial perturbations.

For each strategy, we keep the overall data size constant and implement two variations: we keep 50% and 75% (balanced and light) of the original training data adopted by standard training to balance the training weights, respectively. Robustness is measured using the average contrastive entropy, which includes two components: OOD robustness on BEIR and adversarial robustness on the original MS MARCO dataset. For adversarial robustness evaluation data, we adopt the approach described in Section 3.4; we randomly sample 1,000 queries, and for each query, we select 9 documents at different ranking intervals to generate adversarial samples.

Scaling performance varies from optimization strategies. As shown in Figure 5, we fit the robustness and effectiveness performance of various optimization strategies using model size and data size as contours, respectively. Robustness and effectiveness both tend to improve with increases in model size and data size. Effectiveness is more strongly influenced by model size, while robustness is enhanced by data size. This can be explained by differences in the scaling laws between the two as discussed in Section 3. Different optimization strategies follow similar scaling laws.

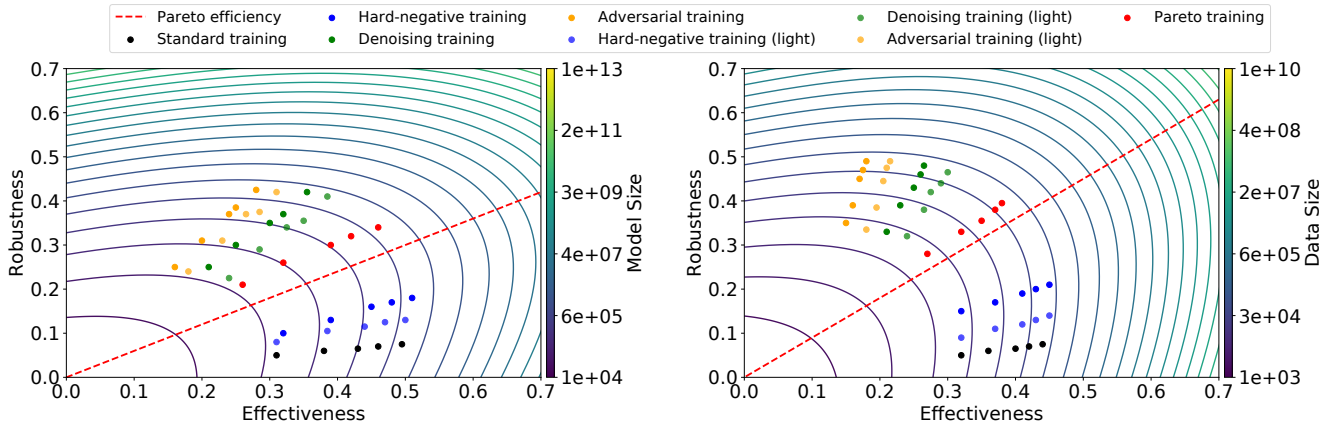


Figure 5: Joint scaling trends in robustness and effectiveness of the BERT model with respect to model size and data size across different optimization strategies. Effectiveness is evaluated on MS MARCO, and robustness is the average of the OOD robustness on BEIR and the adversarial robustness against ranking attacks on MS MARCO. For visualization purposes, the metric in the figure is obtained by inverse normalization of the (average) contrastive entropy.

When comparing optimization strategies along a single contour: (i) Standard training exhibits the worst robustness and limited effectiveness gains. This suggests that focusing solely on effectiveness can lead to overfitting, and a lack of understanding of robustness can, in turn, constrain effectiveness. (ii) Hard-negative training performs better than standard training in both robustness and effectiveness. It introduces challenging and negative cases for comparison in the optimization objective, which improves the model's ability to discriminate relevant cases, thereby enhancing both robustness and effectiveness. (iii) Denoising training achieves improved robustness compared to hard-negative training, but at the cost of reduced effectiveness. Denoising training enhances the model's robustness to irrelevant features by introducing random noise, but it also harms the relevant features of positive samples. (iv) Adversarial training achieves the best robustness but performs poorly in terms of effectiveness. An excessive focus on robustness can compromise effectiveness.

Pareto efficiency exists between robustness and effectiveness. By keeping the model size and data size constant, we can observe the trade-offs between robustness and effectiveness through different optimization strategies: (i) at the ends of the contour lines, both robustness and effectiveness are relatively low; (ii) as the optimization strategy shifts from focusing on only one aspect, both robustness and effectiveness increase; and (iii) as the weights of the optimization strategy approach balance, robustness and effectiveness exhibit Pareto efficiency. At the Pareto efficiency, robustness has reached a proper position and cannot be better off without making effectiveness worse off, and vice versa.

Deviation from Pareto efficiency leads to suboptimal scaling. When analyzing specific weighting ratios between effectiveness and robustness (e.g., the proportion of adversarial samples in adversarial training or the noise ratio in denoising training), we observe the following (see Figure 5). (i) The allocation of weights impacts both the model's current performance and its scaling behavior. (ii) Examining the contours, we find that existing optimization strategies deviate from Pareto efficiency, resulting in suboptimal scaling directions. While Pareto efficiency allows for the most efficient scaling, even at the Pareto point, the joint performance may match the equivalent scaling performance without further scaling.

Identifying Pareto efficiency is crucial for achieving an optimal balance between robustness and effectiveness. It provides a pathway to enhance joint performance while enabling efficient scaling.

4.2 Pareto Training

Building on the concept of Pareto efficiency, we propose Pareto training to develop dense retrieval models that balance robustness and effectiveness while enabling efficient scaling.

Training method. The key idea of the proposed Pareto training method is to dynamically adjust the weights of optimization objectives during training, allowing the joint performance of robustness and effectiveness to progressively approach Pareto efficiency. A major challenge lies in the fact that the training losses for robustness and effectiveness do not always align with the model's final performance, making it difficult to achieve Pareto efficiency during training. To address this, we adopt the concept of *distributionally robust optimization* [8] by dynamically adjusting the loss weights to estimate the gap with Pareto efficiency, thus deriving an approximate solution for the optimal weight ω.

To initialize the weight ω_0 we calculate the ratio between robustness and effectiveness at Pareto efficiency. We jointly optimize the robustness and effectiveness objectives using the following loss:

$$\mathcal{L}_{\text{Pareto}}(\theta) = \omega \ell_R(\theta) + (1-\omega)\ell_E(\theta), \tag{9}$$

where θ represents the parameters of the dense retrieval model, and $\ell_R(\cdot)$ and $\ell_E(\cdot)$ denote the robustness loss and effectiveness loss, respectively. The training loss on both robustness data and effectiveness data is implemented by pairwise loss functions.

At each training step $t+1$ ($t \geq 0$), we dynamically update the weight ω using the formula:

$$\omega^{t+1} = \omega^t - \eta \left(\frac{\ell_E^t(\theta)}{\ell_R^t(\theta)} - \frac{1}{\omega_0} \right), \tag{10}$$

$$\omega^{t+1} = \min(\max(\omega^{t+1}, 0), 1), \tag{11}$$

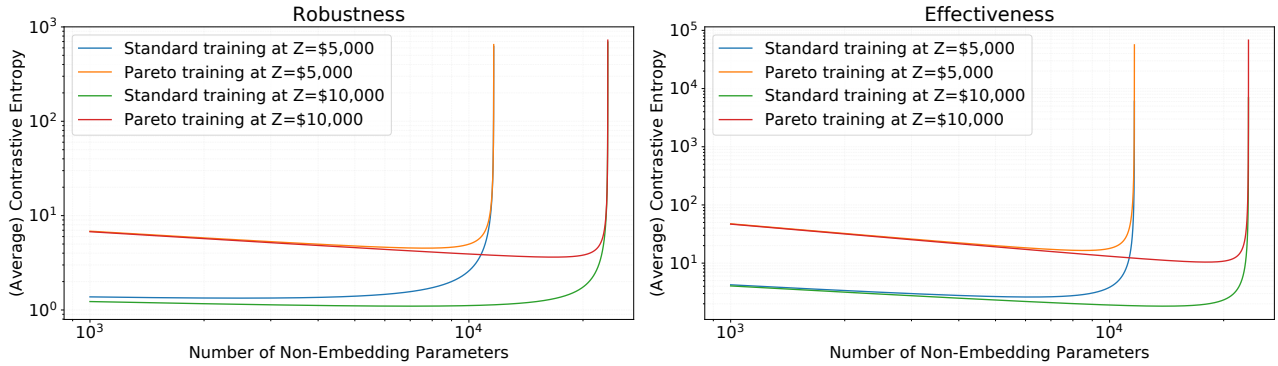


Figure 6: Predicted contrastive entropy of robustness and effectiveness for standard and Pareto training under limited budgets.

where η is the learning rate and set to 0.1.

Experimental setup. Pareto training can be adapted to various optimization strategies. We use adversarial training as an example, as it is somewhat distant from Pareto efficiency. We follow the adversarial training method from Section 3.4 and apply Pareto training to dynamically adjust training weights between adversarial sample optimization and original sample optimization to approach Pareto efficiency. We conduct experiments within the existing range of model and data sizes and observe the joint performance and scaling behavior on English benchmarks. For robustness testing, we continue to combine OOD robustness and adversarial robustness.

Pareto training efficiently optimizes robustness and effectiveness in a joint perspective. In Figure 5 we observe that (i) Pareto training closely approaches Pareto efficiency in both model and data scaling. This indicates that our proposed training method can adaptively adjust the optimization direction, thereby improving the joint performance of robustness and effectiveness. (ii) Pareto training demonstrates a significant improvement in scaling efficiency, indirectly validating the accuracy of the Pareto efficiency we have fitted. (iii) Even without scaling, the performance of the Pareto–trained model is equivalent to the result of large-scale scaling of the standard training model.

5 Resource-Friendly Application

We estimate the resource costs across the lifecycle of a dense retrieval model, including data preparation, model training, and model inference. We discuss, from the joint perspective of robustness and effectiveness, how to deploy resource-friendly dense retrieval models.

5.1 Data-Model Joint Scaling Laws

To study the resource costs, we consider the scaling laws of both data and models jointly. Following [6], the joint effects of model size and data size can be approximated by:

$$L(f, \mathcal{D}_{\text{train}}) = L(f) + L(\mathcal{D}_{\text{train}}) \approx \left[\left(\frac{M}{|f|}\right)^{\frac{\mu}{\eta}} + \frac{D}{|\mathcal{D}_{\text{train}}|}\right]^{\eta} + \delta, \tag{12}$$

where $|f|$ and $|\mathcal{D}_{\text{train}}|$ represent the model size and data size, respectively, and M, D, μ, η, δ are coefficients. This equation can be used to measure the total loss of either robustness or effectiveness. In this section the robustness remains an average of OOD robustness and adversarial robustness; the setting is consistent with the Section 4. We take the English benchmark, approximate the parameters associated with standard training and Pareto training by numerical estimation; see Table 4. Based on the joint scaling law, we can study the relationship between resource costs and performance.

Table 4: Estimated parameters in data-model joint scaling laws on English benchmark for standard and Pareto training.

Training	Aspect	M	D	μ	η	δ
Standard	Robust.	2.11×10^3	2.99×10^3	0.10	0.78	0.01
	Effect.	3.47×10^4	2.14×10^3	0.38	1.10	0.04
	Aspect	M	D	μ	η	δ
Pareto	Robust.	1.96×10^4	2.57×10^3	0.25	0.80	0.02
	Effect.	8.34×10^5	2.17×10^3	0.57	1.38	0.03

5.2 Resource Budget Allocation

Approximately, the total cost of training and inference of a dense retrieval model with a data size of $|\mathcal{D}_{\text{train}}|$ and model size $|f|$ is:

$$Z(f, \mathcal{D}_{\text{train}}) = Z_{\text{data}} \cdot |\mathcal{D}_{\text{train}}| + Z_{\text{train}} \cdot |f| + Z_{\text{infer}} \cdot |f|, \tag{13}$$

$$Z_{\text{data}} \approx 0.6,\ Z_{\text{train}} \approx 3.22 \times 10^{-8},\ Z_{\text{infer}} \approx 0.43, \tag{14}$$

where Z_{data}, Z_{train}, Z_{infer}, represent cost factors corresponding to data preparation, training, and inference, respectively. They are given by prior research [6], with the units in dollars. Notably, in our setup, the data costs Z_{data} not only include manual annotation but also automatically generated data, such as adversarial examples. These costs are relatively lower than manual annotation, so we can treat them as having equivalent costs for the sake of simplicity.

By combining Eq. 12 and 13, we can observe the trend of how robustness and effectiveness change with model size scaling under a fixed budget. We select two budget levels ($Z = \$5,000$ and $Z = \$10,000$) and compare the trends of standard training and Pareto training in terms of robustness and effectiveness. The results are shown in Figure 6. (i) Within a certain range, both methods can achieve improvements in robustness and effectiveness by increasing the model size. However, indiscriminate scaling can lead to a limited available budget for data, which ultimately harms performance. (ii) Due to its relative insensitivity to model scaling, robustness often experiences performance degradation earlier. When allocating the budget, we need to consider both robustness and effectiveness to achieve a balanced performance in dense retrieval models. (iii) By comparison, due to its ability to achieve efficient scaling, Pareto training exhibits a scaling efficiency improvement of about 2.5 times

compared to standard training when the budget is 5,000. It can fully use limited budgets, and scale to larger models to achieve better joint performance in terms of robustness and effectiveness.

6 Related Work

Dense retrieval. As the size of neural networks increases, dense vectors output by the models contain rich information and exhibit strong discriminative power, making them suitable for indexing and distinguishing documents [10, 16, 54]. This has led to the emergence of dense retrieval, which stands out among retrieval model families due to its effectiveness [22]. Subsequent research has focused on enhancing the effectiveness of dense retrieval, such as increasing model capacity [32] and expanding training data [4]. Only optimizing for effectiveness may lead to a loss in robustness.

Robustness is another crucial metric for dense retrieval models. Existing studies have found that despite their remarkable effectiveness, dense retrieval models sometimes fall short in robustness compared to traditional sparse retrieval models [41]. Existing work mainly focuses on OOD robustness and adversarial robustness for dense retrieval. E.g., Thakur et al. [41] and Liu et al. [23] have revealed flaws of dense retrieval models in terms of OOD robustness, while Liu et al. [24] Zhong et al. [55] have identified vulnerabilities to adversarial robustness. Concerns about robustness have hindered the widespread application of dense retrieval models in real-world scenarios, and current research lacks a precise understanding of the scaling laws of robustness.

Both robustness and effectiveness are important for dense retrieval models. Some work claims that robustness and effectiveness are conflicting goals that are difficult to optimize simultaneously [42, 46]. Others have found that models with strong effectiveness also demonstrate good robustness, revealing that there is potential for the two to be co-improved [25, 27]. Nevertheless, existing work lacks quantitative analysis, leaving a limited understanding of their relationship. In this paper, we explicitly address the relationship between robustness and effectiveness by investigating their scaling laws and their trade-off performance.

Scaling laws. Zipf [56] reveals an inverse relationship between word frequency and its rank in the frequency distribution of natural language. Heaps' law [7] describes the growth pattern of vocabulary size with the total number of words in a document, becoming a key principle in estimating inverted indexes in information retrieval. With the development of neural networks, research on scaling laws has gradually shifted to the size of neural models, dataset sizes, and computational resources. E.g., Hestness et al. [11] discover a power-law relationship, which has since been extended to larger models [15] and eventually quantified exactly [12]. These findings provide insights for predicting model performance and rationally allocating training resources. In IR, GTR uses scaling laws to increase model size and enhance dense retrieval performance [32]. Fang et al. [6] precisely fit the scaling laws of dense retrieval effectiveness using formulas. These works primarily focus on the scaling laws of effectiveness, neglecting the changes in model robustness and the relationship between robustness and effectiveness.

7 Conclusion

We have presented a comprehensive analysis of scaling laws for robustness and effectiveness in dense retrieval, exploring methods to jointly optimize both while maintaining efficiency. By scaling model and data sizes, we observe that the robustness of dense retrieval adheres to scaling laws. The scaling patterns of robustness and effectiveness differ, resulting in significant resource overheads for joint optimization. We reveal the existence of Pareto efficiency between robustness and effectiveness, with typical optimization strategies often yielding suboptimal scaling performance due to deviations from this balance. To address this, we propose Pareto training, which achieves an optimal trade-off between robustness and effectiveness, enabling efficient model scaling. Even without scaling, Pareto training consistently enhances both robustness and effectiveness. We also show how Pareto efficiency can guide resource allocation for practical deployment scenarios. We hope that our findings can contribute to the development of resource-friendlier IR.

Limitations and future work. (i) For OOD robustness, the BEIR benchmark is primarily used for evaluation. However, the uneven data distribution across datasets in the benchmark may impact the stability of our results. In future work, more realistic simulations of robustness scenarios warrant further investigation. (ii) While this study focused on OOD robustness and adversarial robustness, which are well-established in information retrieval, future research should explore a broader range of robustness types, including performance variance [52] and retrievability [1]. (iii) We investigate scaling laws for the robustness of dense retrieval models in terms of model size, data size, optimization strategy, and annotation quality, leaving aspects like scaling laws for computational costs [15] and attack costs [13] for future work. (iv) For model architecture, we focus on representative dual-encoder architectures that map queries and documents into embeddings of the same dimension through a shared encoder. Other architectures, such as multi-vector generation [18], hybrid models [39], or interaction-based dense retrieval models [14], may offer different scaling performances and deserve further exploration in future studies. (v) We examine three optimization strategies; methods such as customized pre-training [30], distillation [36], and domain-adaptive training [29] remain unexplored. These strategies could impact model robustness. (vi) Due to time and resource constraints, this study focuses on scaling models within BERT-level pre-trained architectures, with future work planned for validation on large language models.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (NSFC) under Grants No. 62472408, 62372431 and 62441229, the Strategic Priority Research Program of the CAS under Grants No. XDB0680102 and XDB0680301, the National Key Research and Development Program of China under Grants No. 2023YFA1011602, the Youth Innovation Promotion Association CAS under Grants No. 2021100, the Lenovo-CAS Joint Lab Youth Scientist Project, and the project under Grants No. JCKY2022130C039. This work was also (partially) funded by the Dutch Research Council (NWO), under project numbers 024.004.022, NWA.1389.20.183, and KICH3.LTP.20.006, and the European Union's Horizon Europe program under grant agreement No. 101070212.

All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

[1] Leif Azzopardi and Vishwa Vinay. 2008. Retrievability: An Evaluation Measure for Higher Order Information Access Tasks. In *CIKM*. 561–570.

[2] Carlos Castillo and Brian D. Davison. 2011. Adversarial Web Search. *Foundations and Trends in Information Retrieval* 4, 5 (2011), 377–486.

[3] Xuanang Chen, Ben He, Kai Hui, Le Sun, and Yingfei Sun. 2023. Dealing with Textual Noise for Robust and Effective BERT Re-ranking. *IPM* 60 (2023), 103135.

[4] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. 2021. MS Marco: Benchmarking Ranking Models in the Large-Data Regime. In *SIGIR*. 1566–1576.

[5] Zhe Dong, Jianmo Ni, Daniel M Bikel, Enrique Alfonseca, Yuan Wang, Chen Qu, and Imed Zitouni. 2022. Exploring Dual Encoder Architectures for Question Answering. In *EMNLP*. 9414–9419.

[6] Yan Fang, Jingtao Zhan, Qingyao Ai, Jiaxin Mao, Weihang Su, Jia Chen, and Yiqun Liu. 2024. Scaling Laws for Dense Retrieval. In *SIGIR*. 1339–1349.

[7] Alexander Gelbukh and Grigori Sidorov. 2001. Zipf and Heaps Laws' Coefficients Depend on Language. In *CICLing*. Springer, 332–335.

[8] Joel Goh and Melvyn Sim. 2010. Distributionally Robust Optimization and Its Tractable Approximations. *Operations Research* 58, 4-part-1 (2010), 902–917.

[9] Tejas Gokhale, Swaroop Mishra, Man Luo, Bhavdeep Sachdeva, and Chitta Baral. [n. d.]. Generalized but not Robust? Comparing the Effects of Data Modification Methods on Out-of-Domain Generalization and Adversarial Robustness. In *ACL*. 2705–2718.

[10] Jiafeng Guo, Yinqiong Cai, Yixing Fan, Fei Sun, Ruqing Zhang, and Xueqi Cheng. 2022. Semantic Models for the First-stage Retrieval: A Comprehensive Review. *TOIS* 40, 4 (2022), 1–42.

[11] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. 2017. Deep Learning Scaling Is Predictable, Empirically. *arXiv preprint 1712.00409* (2017).

[12] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. An Empirical Analysis of Compute-Optimal Large Language Model Training. *NIPS* 35 (2022), 30016–30030.

[13] Nikolaus Howe, Ian McKenzie, Oskar Hollinsworth, Michał Zajac, Tom Tseng, Aaron Tucker, Pierre-Luc Bacon, and Adam Gleave. 2024. Effects of Scale on Language Model Robustness. *arXiv preprint arXiv:2407.18213* (2024).

[14] Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring. In *ICLR*.

[15] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361* (2020).

[16] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *EMNLP*. 6769–6781.

[17] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

[18] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *SIGIR*. 39–48.

[19] Oren Kurland and Moshe Tennenholtz. 2022. Competitive Search. In *SIGIR*. 2838–2849.

[20] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In *ACL*. 6086–6096.

[21] Chaofan Li, Zheng Liu, Shitao Xiao, Yingxia Shao, and Defu Lian. 2024. Llama2Vec: Unsupervised Adaptation of Large Language Models for Dense Retrieval. In *ACL*. 3490–3500.

[22] Jimmy Lin. 2022. A Proposed Conceptual Framework for a Representational Approach to Information Retrieval. In *ACM SIGIR Forum*, Vol. 55.

[23] Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Wei Chen, and Xueqi Cheng. 2023. On the Robustness of Generative Retrieval Models. In *Gen-IR@SIGIR*.

[24] Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng. 2023. Black-Box Adversarial Attacks against Dense Retrieval Models: A Multi-View Contrastive Learning Method. In *CIKM*. 1647–1656.

[25] Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Multi-granular Adversarial Attacks against Black-box Neural Ranking Models. In *SIGIR*. 1391–1400.

[26] Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Robust Neural Information Retrieval: An Adversarial and Out-of-distribution Perspective. *arXiv preprint arXiv:2407.06992* (2024).

[27] Yu-An Liu, Ruqing Zhang, Mingkun Zhang, Wei Chen, Maarten de Rijke, Jiafeng Guo, and Xueqi Cheng. 2024. Perturbation-Invariant Adversarial Training for Neural Ranking Models. In *AAAI*, Vol. 38. 8832–8840.

[28] Dingkun Long, Qiong Gao, Kuan Zou, Guangwei Xu, Pengjun Xie, Ruijie Guo, Jian Xu, Guanjun Jiang, Luxi Xing, and Ping Yang. 2022. Multi-CPR: A Multi Domain Chinese Dataset for Passage Retrieval. In *SIGIR*. 3046–3056.

[29] Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2021. Zero-shot Neural Passage Retrieval via Domain-targeted Synthetic Question Generation. In *ECIR*. 1075–1088.

[30] Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, Xiang Ji, and Xueqi Cheng. 2021. Prop: Pre-training with Representative Words Prediction for Ad-hoc Retrieval. In *WSDM*. 283–291.

[31] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated Machine Reading Comprehension Dataset. In *CoCo@NIPS*.

[32] Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, et al. 2022. Large Dual Encoders Are Generalizable Retrievers. In *EMNLP*. 9844–9855.

[33] Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document Expansion by Query Prediction. *arXiv preprint arXiv:1904.08375* (2019).

[34] OpenAI. 2022. Introducing ChatGPT. https://openai.com/blog/chatgpt.

[35] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2021. KILT: a Benchmark for Knowledge Intensive Language Tasks. In *NAACL*. 2523–2544.

[36] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering. In *NAACL*. Association for Computational Linguistics.

[37] Nils Reimers and Iryna Gurevych. 2021. The Curse of Dense Low-Dimensional Information Retrieval for Large Index Sizes. In *ACL*. 605–611.

[38] Stephen E Robertson and Steve Walker. 1994. Some Simple Effective Approximations to the 2-poisson Model for Probabilistic Weighted Retrieval. In *SIGIR'94*. Springer, 232–241.

[39] Tao Shen, Xiubo Geng, Chongyang Tao, Can Xu, Guodong Long, Kai Zhang, and Daxin Jiang. 2023. Unifier: A Unified Retriever for Large-scale Retrieval. In *SIGKDD*. 4787–4799.

[40] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: Enhanced Representation Through Knowledge Integration. *arXiv preprint 1904.09223* (2019).

[41] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *NIPS*.

[42] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2019. Robustness May Be at Odds with Accuracy. In *ICLR*.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NIPS*, Vol. 30.

[44] Joel Watson. 2013. *Strategy: An Introduction to Game Theory* (3 ed.). W. W. Norton and Company.

[45] Chen Wu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2023. PRADA: Practical Black-Box Adversarial Attacks against Neural Ranking Models. *TOIS* 41, 4 (2023), Article 89.

[46] Chen Wu, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022. Are Neural Ranking Models Robust? *TOIS* 41, 2 (2022), 1–36.

[47] Zhijing Wu, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Investigating Passage-Level Relevance and Its Role in Document-Level Relevance Judgment. In *SIGIR*. 605–614.

[48] Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. 2015. Learning Maximal Marginal Relevance Model via Directly Optimizing Diversity Evaluation Measures. In *SIGIR*. 113–122.

[49] Xiaohui Xie, Qian Dong, Bingning Wang, Feiyang Lv, Ting Yao, Weinan Gan, Zhijing Wu, Xiangsheng Li, Haitao Li, Yiqun Liu, et al. 2023. T2ranking: A Large-Scale Chinese Benchmark for Passage Ranking. In *SIGIR*. 2681–2690.

[50] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *ICLR*.

[51] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. GLM-130B: An Open Bilingual Pre-trained Model. In *ICLR*.

[52] Peng Zhang, Dawei Song, Jun Wang, and Yuexian Hou. 2013. Bias-Variance Decomposition of IR Evaluation. In *SIGIR*. 1021–1024.

[53] Sheng Zhang, Xin Zhang, Hui Wang, Lixiang Guo, and Shanshan Liu. 2018. Multi-Scale Attentive Interaction Networks for Chinese Medical Question Answer Selection. *IEEE Access* 6 (2018), 74061–74071.

[54] Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2022. Dense Text Retrieval based on Pretrained Language Models: A Survey. *arXiv preprint arXiv:2211.14876* (2022).

[55] Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. 2023. Poisoning Retrieval Corpora by Injecting Adversarial Passages. In *EMNLP*.

[56] George Kingsley Zipf. 2016. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Ravenio Books.

[57] Xin Zou and Weiwei Liu. 2024. On the Adversarial Robustness of Out-of-Distribution Generalization Models. *NIPS* 36 (2024).