Scaling Laws For Dense Retrieval

Yan Fang*
fangy21@mails.tsinghua.edu.cn
Department of Computer Science and Technology, Tsinghua University
Quan Cheng Laboratory
Beijing 100084, China

Jingtao Zhan*
jingtaozhan@gmail.com
Department of Computer Science and Technology, Tsinghua University
Quan Cheng Laboratory
Beijing 100084, China

Qingyao Ai†
aiqy@tsinghua.edu.cn
Quan Cheng Laboratory
Department of Computer Science and Technology, Tsinghua University
Beijing 100084, China

Jiaxin Mao
maojiaxin@gmail.com
Gaoling School of Artificial Intelligence, Renmin University of China
Beijing 100872, China

Weihang Su
swh22@mails.tsinghua.edu.cn
Department of Computer Science and Technology, Tsinghua University
Zhongguancun Laboratory
Beijing 100084, China

Jia Chen
chenjia2@xiaohongshu.com
Xiaohongshu Inc
Beijing, China

Yiqun Liu
yiqunliu@tsinghua.edu.cn
Department of Computer Science and Technology, Tsinghua University
Zhongguancun Laboratory
Beijing 100084, China

ABSTRACT

Scaling laws have been observed in a wide range of tasks, particularly in language generation. Previous studies have found that the performance of large language models adheres to predictable patterns with respect to the size of models and datasets. This helps us design training strategies effectively and efficiently, especially as large-scale training becomes increasingly resource-intensive. Yet, in dense retrieval, such scaling law has not been fully explored. In this study, we investigate how scaling affects the performance of dense retrieval models. We implement dense retrieval models with different numbers of parameters, and train them with various amounts of annotated data. We propose to use the contrastive entropy as the evaluation metric, which is continuous compared with discrete ranking metrics and thus can accurately reflect model performance. Results indicate that the performance of dense retrieval models follows a precise power-law scaling related to the model size and the number of annotations across different datasets and annotation methods. Additionally, we show that the scaling laws help optimize the training process, such as resolving the resource allocation problem under a budget constraint. We believe that these findings significantly contribute to understanding the scaling effect of dense retrieval models and offer meaningful guidance for future research.

*Both authors contributed equally to this research.
†Corresponding author

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657743

CCS CONCEPTS

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

KEYWORDS

Dense retrieval, Neural scaling law, Large language models

ACM Reference Format:
Yan Fang, Jingtao Zhan, Qingyao Ai, Jiaxin Mao, Weihang Su, Jia Chen, and Yiqun Liu. 2024. Scaling Laws For Dense Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3626772.3657743

1 INTRODUCTION

The studies of scaling laws in language data can be traced back to a century ago. In the 1920s, a couple of linguisticians discovered that the frequency of a word is proportional to the inverse of its rank when sorting vocabulary based on each word's frequency in the corpus, which is widely known as the Zipf's law [2, 33]. Later in the 1960s, Gustav Herdan found that the number of distinct words in a corpus approximately follows a function of the corpus size, which can be approximated with a power function. This is often referred to as the Heaps's law [27]. These foundational discoveries in scaling laws have profoundly influenced research in linguistics and information retrieval. For example, Zipf's law has inspired the development of several statistical retrieval models, and Heap's law has served as the key principle for the estimation of inverted index, the foundation of many retrieval systems.

Recently, as language modeling has evolved from statistical analysis to the learning of semantic representations, the focus of scaling law research has also shifted from analyzing text statistics toward the training dynamics of large language models (LLMs). Significant

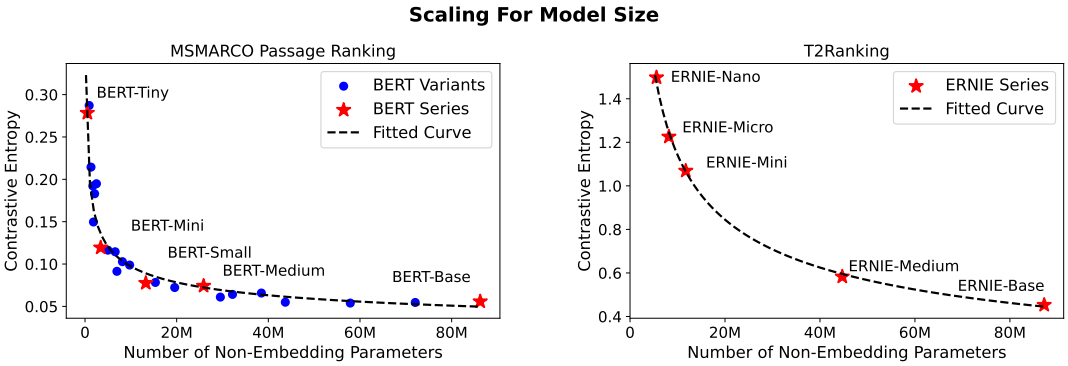

Figure 1: Performance of various models on MSMARCO Passage Ranking (left) and T2Ranking (right) datasets. It shows the number of non-embedding parameters (x-axis) and the test-set contrastive entropy (y-axis). The stars and points represent the actual performance. The curves are derived from the scaling law and match the observed data.

research effort has been dedicated to examining how different factors influence model performance, such as model size, data volume, and computational capacity [22]. Findings from these studies reveal precise power law relationships between model performance and scaling factors, which enable researchers and developers to empirically predict model performance without actually constructing the models [1]. Since the training of modern LLMs demands substantial time and financial resources, such scaling laws are of great importance in practice.

Similar to language modeling, Dense Retrieval models have emerged as a significant milestone in this transition from statistical analysis to semantic representation learning in Information Retrieval [5, 26]. In contrast to traditional statistical retrieval methods such as BM25 [43], Dense Retrieval models are initialized with pre-trained language models and finetuned on annotation data in an end-to-end manner. They capture the semantic similarity between queries and documents, and demonstrate superior performance over traditional methods [31, 39, 46]. However, researchers find that the effectiveness of dense retrieval models is sensitive to multiple training factors [53, 56]. Therefore, the construction of effective dense retrieval models under practical constraints (such as budget and latency requirements) is not straightforward, and more insights on the optimization process of Dense Retrieval are needed.

In this paper, we investigate the scaling laws for dense retrieval models[1]. While some studies have indicated that larger models exhibit improved generalization capabilities in zero-shot dense retrieval tasks [35, 44], to the best of our knowledge, there isn't any published literature explicitly discover scaling laws in dense retrieval models. Specifically, there are two challenges; (1) traditional performance metrics in retrieval tasks (e.g., NDCG) are discrete functions, which limits their ability to stably and smoothly reflect the change of model performance in practice; (2) the training process of Dense Retrieval involve multiple interrelated factors such as model size, annotation size, and annotation quality, which makes it difficult to isolate the effect of each factor separately. To this end, we first propose to evaluate the quality of dense retrieval models with a contrastive entropy metric. The idea is inspired by the popular contrastive ranking loss and the analysis of token generation perplexity in LLMs. It measures the likelihood of retrieving a relevant document from a randomly sampled candidate set, and shares a similar structure with the training loss of dense retrieval models. The smooth nature of this metric considerably facilitates our subsequent analysis. Second, to disentangle the effects of model size and data size in dense retrieval, we conducted experiments with models implemented with different pre-trained language models with non-embedding parameter sizes ranging from 0.5 to 87 million, on two of the largest web search datasets, i.e., MSMARCO and T2Ranking. Experimental results show that, under proper experimental conditions, the performance of dense retrieval models follows a precise power-law scaling with respect to training factors. Figure 1 illustrates such power-law scaling with model size. To investigate the effect of annotation quality, we adopted several LLMs and weak supervision methods to generate training data for dense retrieval models. Our results indicate that the observed scaling laws of dense retrieval are uniformly valid across models trained with different types of annotation data. Additionally, we show that the joint effect of model and data sizes can be nicely fitted and predicted with a single function within a certain range. Such functions can be used to find the best resource allocation strategy given a restricted budget, and could potentially provide important insights for the practical implementation of dense retrieval models and green IR [47].

This paper is organized as follows. We first breifly revisit the related work in Section 2. Then, we present our systematic evaluation framework in Section 3. With this framework, we investigate the scaling laws of dense retrieval in Section 4 and show its potential application in Section 5.

2 BACKGROUND AND RELATED WORK

In this section, we revisit the background about scaling laws and dense retrieval. We start with the scaling laws in linguistic analysis and in neural language models. Then we present the explorations about dense retrieval techniques and its training technique.

[1]Code is open-sourced at https://github.com/jingtaozhan/DRScale.

2.1 Scaling Laws in Linguistic Language Data

Zipf's law [2, 33] is a well-known evidence about the existence of universal power laws in cognitive science and the social sciences. It shows an inverse correlation between the frequency of a word's occurrence in natural language and its rank in the frequency distribution. It is widely applied in different areas. Furthermore, Zipf's law is tightly connected to other statistical scaling laws in linguistics, notably Heaps' law [16, 25, 27]. Heaps' law shows a sublinear growth trajectory between a text's vocabulary size and its total word count. As the total word count increases, the rate of introducing new words diminishes, leading to a plateau in vocabulary expansion. This phenomenon is particularly significant in information retrieval, which serves as the key principle for the estimation of inverted index.

2.2 Neural Scaling Law

Neural scaling law describes the relationship between model size, dataset size, computational budget, and performance in neural network training. This concept was first introduced by Hestness et al. [17] as a power-law relationship. Subsequently, Kaplan et al. [22] expanded it to larger models. Hoffmann et al. [18] further refined it by developing a unified formula for scaling laws, incorporating data-dependent scaling terms for compute-optimal training.

These empirical scaling laws offer crucial insights for training large Transformer-based models, particularly by accurately predicting loss. Notably, experimental results from smaller models can be extrapolated to larger ones. Recent studies show that such scaling laws also hold for many other model architectures. For instance, Clark et al. [6] investigated the scaling laws in Mixture of Experts (MoE) models. Gao et al. [15] showed the scaling effects in model optimization with Reinforcement Learning.

Beyond language-centric tasks, these scaling principles have been adapted for domain-specific applications, such as speech recognition [41], computer vision [8, 55], and multi-modal language-vision settings [21, 38, 40]. In Information Retrieval (IR), Ardalani et al. [4] investigated the application of scaling laws in Click-Through Rate (CTR) recommendation tasks, and Zhang et al. [57] addressed their relevance in conventional ID-based sequential recommendation models. Nonetheless, there has been limited research into whether scaling laws remain applicable in dense retrieval.

2.3 Dense Retrieval

We now briefly revisit prior studies in the field of dense retrieval. The training data for dense retrieval tasks typically comprises annotated pairs, each consisting of a query and a human-labeled relevant passage. Early research primarily concentrated on effective negative sampling strategies used for dense retrieval training, such as employing random passages or the top irrelevant passages retrieved by BM25 as negative samples [23]. ANCE [53] utilized self-mined hard negatives and substantially improved the retrieval performance. Furthermore, Zhan et al. [56] proposed dynamic hard negatives to further enhance both training efficiency and retrieval effectiveness. RocketQA [39] and TAS-B [19] introduced knowledge distillation, utilizing a well-trained cross-encoder model to generate soft labels for training pairs.

Beyond the design of finetuning methods, researchers also explore other techniques, such as pretraining methods and multi-vector retrieval. (1) Pretraining studies design objectives that are similar to the retrieval tasks. For example, Condenser [13] and coCondenser [14] use the Sequence Contrastive Learning task to improve the representational capability. RetroMAE [51] leverages an encoder-decoder architecture, wherein a shallow decoder encourages the encoder to produce higher-quality representations. Contriever [20] pre-trains dense retrieval models with Inverse Cloze Task and the Independent Cropping Task. (2) Since the single vector representation in dense retrieval could become a limitation, various studies have explored more complex scoring techniques. ME-BERT [28] introduces multi-vector representations to enable more precise retrieval of long documents. ColBERT [24, 45] investigates token-level vector representations and aggregates scores using a late-interaction mechanism. Other researchers attempt to expand the vector dimension to vocabulary size [11, 12]. This expansion allows dense retrieval models to directly generate term weights, facilitating retrieval similar to sparse models.

Prior explorations of dense retrieval models mainly focus on techniques with a static setup, such as a certain model size, certain data size, etc. Instead, we employ a dynamic setup and explore how model perform when the model size and data size are scaled.

2.4 Query Generation

Besides human-labeled data, dense retrieval can also utilize query generation techniques to generate pseudo annotations [29, 49]. Query generation involves generating multiple relevant queries for a given passage [36, 37]. The most basic approach employs unsupervised heuristic methods, such as the previously mentioned Sequence Contrastive Learning (SCL) or Inverse Cloze Task (ICT). However, the quality of the weak supervision data generated by these methods is relatively low. Therefore, they are primarily used in the unsupervised pre-training phase due to their accessibility. More advanced methods leverage pre-trained language models like T5 to generate more precise relevant queries for data augmentation [36]. Nevertheless, these generated queries are often used for document expansion to enhance the retrieval performance in lexical matching models. As training data, these queries are usually exploited in scenarios where human annotations are scarce, such as in out-of-domain situations.

3 METHEDOLOGY

In this section, we first introduce the model architecture and datasets used for exploring the scaling effect of dense retrieval. We further discuss the training strategy used in the experiments and the proposed performance evaluation metrics.

3.1 Problem Formulation

We first formalize the dense retrieval model. For a given corpus, the goal is to identify the top relevant passages for a specific query. Dense retrieval models accomplish this by employing an encoder that maps both queries and candidate passages into a shared dense embedding space. Subsequently, a scoring function, such as inner product or cosine similarity, is applied to the encoded dense vectors to compute relevance scores. Let q and p be the query and the

passage, respectively. We use $f(\cdot;\theta)$ to denote the the mapping function of the dense retrieval model parameterized by θ. The relevance score $s(q,p)$ is as follows:

$$s(q,p) = \langle f(q;\theta), f(p;\theta)\rangle \tag{1}$$

In this paper, we only consider that the encoders for queries and passages are shared, as it is a popular implementation choice in practice. We leave the studies of separate query and document encoders to future studies.

The training data for dense retrieval typically comprises a set of training queries and associated human annotations. Each query is annotated with one or more relevant passages, and the remaining unannotated passages are generally presumed irrelevant. In this paper, we adhere to this annotation standard and consider each query-positive-passage pair as an individual data point. Formally, the training set consists of n data points, $\{(q_i, p_i^+)\}_{i=1}^n$, where q_i and p_i^+ denote the i-th query in the training set and its corresponding annotated positive passage.

3.2 Model Architechture

With the development of large-scale pre-trained language models, advanced dense retrieval models in recent years have followed the Transformer's structure. While some studies have explored using decoder-only architectures to generate dense vector representations of texts, mainstream dense retrieval models still employ encoder-only models such as BERT due to its bi-directional modeling ability. Formally, a pre-trained Transformer, augmented with a projection layer, serves as the text encoder:

$$v = (\text{Transformer}(x))\, W + b \tag{2}$$

where x represents the text input, and W and b are the parameters of the projection layer.

Typically, the generated vector representation is derived from the [CLS] token representation (in BERT series models) or the mean pooling of the outputs from the last Transformer layer. The main function of the projection layer is to map these vectors into the target semantic space.

In our study, we experimented with Transformer models of various model sizes. With limited annotated query-passage pairs, it is usually difficult to train a large dense retrieval model from scratch. As a result, most dense retrieval models are initialized with pre-trained language models and then perform fine-tuning on the annotated data. Therefore, to align with prevailing research practices, we focus our analysis on dense retrieval models initialized from different sizes of pre-trained language models.

Previous studies have shown that different pre-training tasks significantly affect the performance of dense retrieval models [13, 14, 20, 30, 51]. To minimize such influence, we select a series of models with identical pre-training configurations and only differ in parameter sizes. Specifically, for experiments on the English corpus, we chose 24 BERT checkpoints from the original Google release [9], with model sizes ranging from 0.5 million (BERT-Tiny) to 82 million parameters (BERT-Base)[2]. For experiments on Chinese retrieval benchmarks, we selected the ERNIE series [48], which were pre-trained on Chinese corpora using tasks similar to BERT. To each model, we attach a projection layer, as shown in Eq. (2), to map the output dimensionality of embeddings to 768 for consistent comparisons.

[2]https://github.com/google-research/bert. Following Kaplan et al. [22], we define the model size as the number of non-embedding parameters.

3.3 Training Data

We utilize publicly available retrieval datasets for exploring the scaling effect for dense retrieval models. To ensure the generalizability and completeness of our study, we follow recent DR research and use MS MARCO Passage Ranking dataset [34] (English) and T2Ranking [52] (Chinese) for the experiments. MS MARCO Passage Ranking is a large-scale annotated dataset with a corpus of 8.8M passages from English web pages and 0.5M training queries. Each training query is coupled with a manually labeled positive passage, which together constitute the annotated pairs. MS MARCO also provides around 7,000 validation queries for performance evaluation. T2Ranking is a recently released large-scale Chinese benchmark for passage ranking, which comprises more than 300k queries and over 2M unique passages collected from real-world search engines.

3.4 Training Setting

As discussed previously, in this paper, we construct dense retrieval models from the pre-trained language model checkpoints and perform fine-tuning with the annotated query-document pairs in each dataset. One of the most important parts of dense retrieval model training is the negative sampling strategy. Previous work has shown that mining hard negative samples in the training process can significantly improve the retrieval performance. However, the primary objective of this work is to investigate the scaling effects of dense retrieval models. As a result, we do not focus on sophisticated training strategies. For simplicity, we adopt the most straightforward approaches, namely random negative sampling and in-batch negative techniques, for the training of all dense retrieval models in this paper. These methods are employed to minimize the influence of sampling strategies.

Formally, for each query-passage pair (q_i, p_i^+), we randomly select a set of unlabeled passages from the corpus as the negative. Then we can optimize the following contrastive ranking loss:

$$\mathcal{L}(\theta) = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(s(q_i, p_i^+;\theta)\right)}{\exp\left(s(q_i, p_i^+;\theta)\right) + \sum_j \exp\left(s(q_i, p_j^-;\theta)\right)} \tag{3}$$

where B denotes the training batch size, $\{p_j^-\}$ is the set of negative passages and $s(q,p;\theta)$ is the scoring function of query and passage:

$$s(q,d;\theta) = \langle f(q;\theta), f(d;\theta)\rangle \tag{4}$$

Here, $\langle\cdot\rangle$ denotes inner product and θ denotes the parameters of the text encoder.

We fine-tune the models for a fixed 10,000 steps and random sample 256 negatives at each step.

3.5 Evaluation Protocol

We now discuss how we evaluate the retrieval performance. The most widely adopted retrieval paradigm is to rank passages in the corpus based on the relevance scores predicted by the retrieval model and retrieve the Top-K candidates to form a ranked list. The performance of the retrieval model is then assessed based on the ranked list using well-defined ranking metrics such as NDCG@K

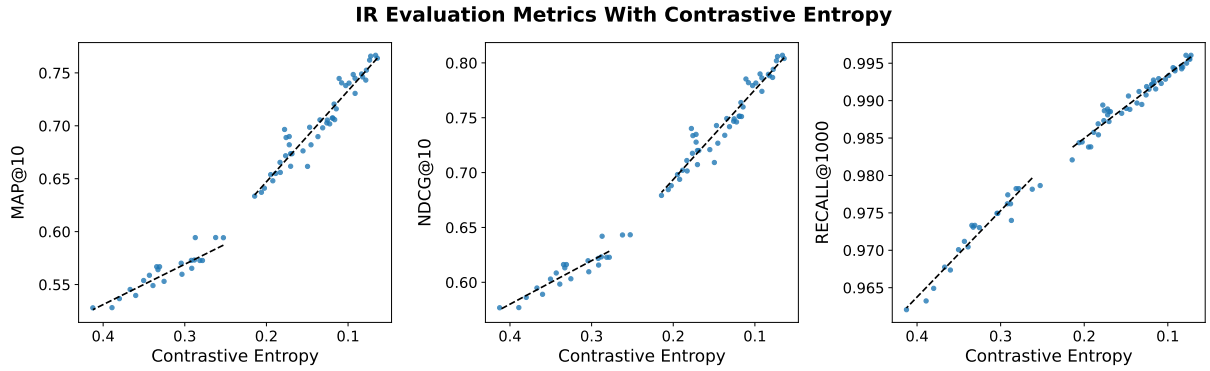

Figure 2: Relationship between standard ranking metrics and contrastive entropy for different Dense Retrieval models on the MSMARCO Passage Ranking dataset. The figures illustrate the contrastive entropy (x-axis) versus standard ranking metrics (y-axis). The results indicate a strong positive correlation. Besides, the figures highlight an emergent ability phenomenon [50] around a contrastive entropy value of approximately 0.25, where there is a significant improvement in ranking metrics.

and MAP@K. However, such metrics are not continuous due to their discrete nature and reliance on a cutoff parameter, K. Because the ranking metrics of a ranked list would not change unless the sequence of the passages changes, these ranking metrics are not sensitive to the changes of model outputs in many cases. Also, with the cutoff in ranking metric, a positive passage only contributes to the metric when ranked within the top K results. If it falls beyond K, whether at K+1 or further, it has no impact on the metric score. The characteristics of these existing ranking metrics make them unsuitable for the investigation of scaling laws in dense retrieval.

To solve these problems, we propose to utilize a continuous metric that sensitively reflects the overall retrieval capability of the models. Inspired by the analysis of scaling laws in large language models, which utilize the perplexity of token generations as evaluation metrics, we propose to use the contrastive entropy as our evaluation metric. Formally, for each query-passage pair in the test set, we randomly select a fixed number (256 in this paper) of negative passages and define the contrastive entropy as:

$$-\log \frac{\exp\left(s(q_i, p_i^+; \theta)\right)}{\exp\left(s(q_i, p_i^+; \theta)\right) + \sum_j \exp\left(s(q_i, p_j^-; \theta)\right)} \tag{5}$$

We investiagte the correlation between the contrastive entropy and existing ranking metrics. We train multiple dense retrieval models. To efficiently evaluate their retrieval performance, we sample a subset corpus that contains 100,000 passages during evaluation. Figure 2 shows the contrastive entropy and ranking metrics, including MAP@10, NDCG@10, and Recall@1000. We can see that the correlation between the contrastive entropy and existing ranking metrics is strong and positive. It is close to a linear correlation. Therefore, we believe that using contrastive entropy is an effective measure to assess the overall retrieval ability of models in our study.

Figure 2 also shows a critical point around 0.25 contrastive entropy, where the top ranking performance evaluated with traditional metrics substantially improves. We attribute this phenomenon to emergent ranking ability. Concurrently, Du et al. [10] also observe this phenomenon in generation tasks. They find emergent abilities are tightly related to a certain loss value. We leave further exploration to future studies.

4 SCALING LAWS FOR DENSE RETRIEVAL

In this section, we show the results of our experiments and summarize our initial investigation of the scaling laws for dense retrieval. Specifically, we aim to thoroughly investigate the following three research questions:

- How does model size impact dense retrieval performance?
- How does annotated training data size influence dense retrieval performance?
- Do different types of data annotations result in distinct scaling effects on dense retrieval models?

4.1 Model Size Scaling

We finetune models of various sizes using the human-annotated training pairs. The finetuning is performed on the entire training sets. We do not utilize early stopping and instead report the best test set loss throughout the training process. This is mainly to mitigate the influence of suboptimal early stopping, which could lead to models being underfitted or overfitted.

Figure 3 illustrates the contrastive entropy on the test set with respect to model sizes. As shown in the figure, the retrieval performance improves (indicated by a lower test loss) as the model size increases. On the left side of the diagram, red stars represent the official checkpoints of variously sized BERT models, while blue points denote other official variants released concurrently. These variants differ in aspects such as the number of attention heads or feed-forward dimensions. The right diagram, in contrast, only features red stars, as the different shape variants of ERNIE are not publicly available.

Based on the observation, we propose to fit the scaling law in terms of model sizes as follows:

$$L(N) = \left(\frac{A}{N}\right)^{\alpha} + \delta_N \tag{6}$$

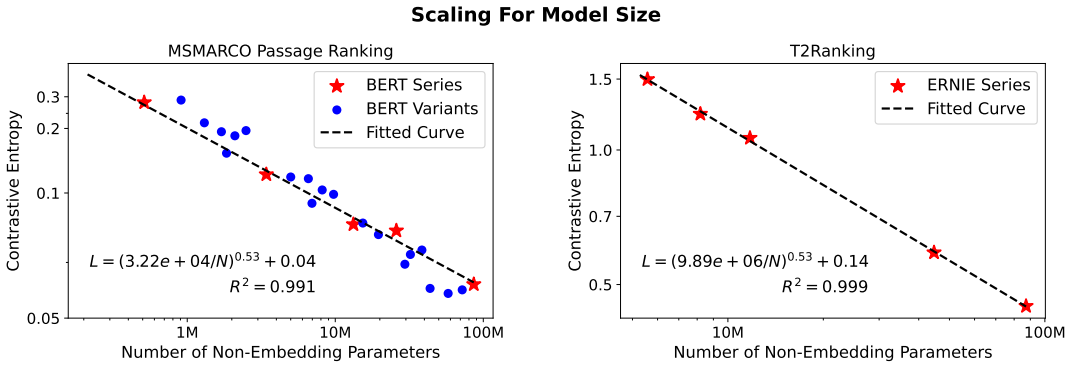

Figure 3: Scaling laws for model performance as a function of model size on MSMARCO Passage Ranking (left) and T2Ranking (right) datasets. The figures display the contrastive entropy (y-axis) against the number of non-embedding parameters (x-axis, logarithmic scale) for different models. Points and stars represent the actual performance, aligning closely along a straight line. The dashed lines are fitted using Eq. (6), demonstrating a close match with the empirical data.

Table 1: Fitting parameters for model size scaling

Dataset	A	α	δ_N	R^2
MSMARCO	3.22×10^4	0.53	0.04	0.991
T2Ranking	9.89×10^6	0.53	0.14	0.999

Table 2: Fitting parameters for data size scaling

Dataset	B	β	δ_D	R^2
MSMARCO	3.49×10^3	1.05	0.05	0.954
T2Ranking	6.04×10^4	0.50	0.15	0.991

where N represents the number of non-embedding parameters of the model, and $L(N)$ denotes the model's contrastive entropy on the test set. Parameters A, α and δ_N are the coefficients.

Note that we introduce a parameter δ_N, which represents a irreducible loss term. It means that a sufficiently large model (setting N to infinity) can only reduces the loss to δ_N rather than zero. This irreducible loss is reasonable given the incomplete annotations and subjective understanding of relevance. On one hand, some relevant passages may not be annotated because they are not sucessfully recalled and are outside the annotation pool. On the other hand, relevance may be subjective to different annotators, which results in even imperfect agreement among different human annotators. Consequently, it is hard for models to perfectly agree with human annotations. Therefore, we believe there should be a irreducible term in the scaling law.

We employ least squares method to fit the linear curve. The coefficients are detailed in Table 1. The coefficient of determination (R^2) suggests a good fit. Based on these results, we validate that the contrastive entropy follows a power-law scaling in relation to the size of non-embedding parameters.

Such discoveries offer new perspectives for future research experiments. For example, given this scaling law, we can initially train smaller models, fit the corresponding scaling curves, and then extrapolate them to predict the performance of larger models. This significantly reduces the cost of conducting experiments directly on larger models and instead offers the opportunity to experiment with different training strategies on smaller models to validate the effectiveness of new approaches.

4.2 Data Size Scaling

We then fix the model size and vary the size of the training data, defined by the number of annotated query-passage pairs. To minimize potential underfit problems caused by small models, we finetune the largest model in this experiment, i.e., the BERT-Base model. Here we present the experiment results up to using all available annotation data.

The results are shown in Figure 4. Similarly, we fit the scaling law in terms of data size with the following log-linear curve:

$$L(D) = \left(\frac{B}{D}\right)^{\beta} + \delta_D \tag{7}$$

where D represents the number of annotated query-passage pairs, and $L(D)$ denotes the contrastive entropy. B, β and δ_D are coefficient to be estimated. The coefficient of determination (R^2) indicates a good fit. Based on these results, we infer that the contrastive entropy follows a power-law scaling relative to the number of annotated query-passage pairs, with specific parameters detailed in Table 2.

This finding offers an alternative perspective for future annotation process. For instance, to determine the amount of annotations for a new corpus, the traditional approach relies on past experience without a clear understanding of the sufficiency of data annotation. With the data-size scaling law, a potential approach is initiating with a minimal amount of annotations, training a model, and fitting the corresponding scaling curve. Accordingly, we can approximate the necessary size of data annotation based on the target performance of the dense retrieval model. This approach establishes a clear relationship between data annotation and the desired performance outcomes. It allows researchers to have a precise expectation

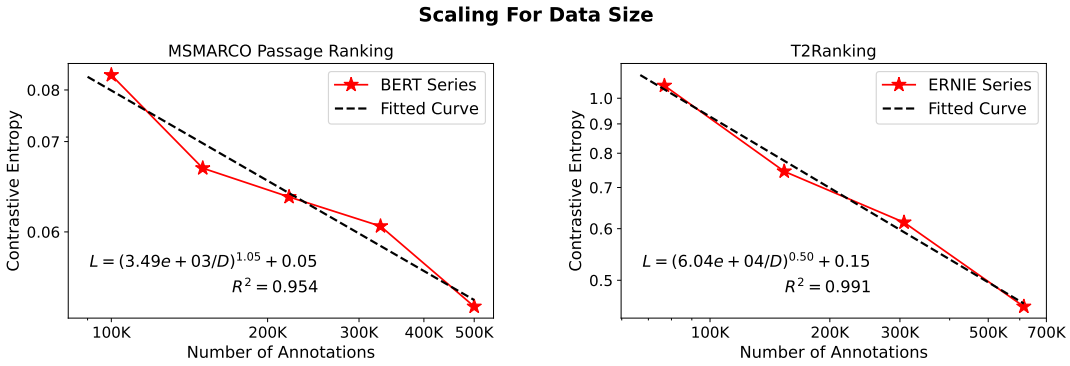

Figure 4: Scaling laws for model performance relative to training data size on MSMARCO Passage Ranking (left) and T2Ranking (right) datasets. The figures illustrate the contrastive entropy (y-axis) as a function of the number of annotated query-passage pairs (x-axis, logarithmic scale) for a fixed model size. Points and stars show the actual performance, aligning closely with a straight line. The dashed lines are fitted using Eq. (7), demonstrating a strong fit with the empirical data.

of future model performance, facilitating more effective planning and budgeting for annotation tasks.

4.3 Annotation Quality

So far, we have observed strong scaling phenomena of dense retrieval model performance with respect to model sizes and data sizes. Yet, in the IR scenario, another aspect that remained unexplored is the quality of data annotations: *Does the scaling effect hold true for data of different quality?*

To investigate this, we conduct experiments using annotations of different quality. Due to constraints in time and resources, our experiments are exclusively conducted on the MSMARCO Passage Ranking dataset. We employ query generation techniques to create three distinct types of annotations:

- **Inverse Cloze Task (ICT):** ICT extracts sentences from passages and and uses the sentence as pseudo-query for the passage. Since it ignores the semantic information, the generated data is of low quality.
- **Supervised Generation Models:** We utilize docT5query [36] to produce multiple queries for each passage. DocT5query is trained on human annotations. The generated data is of higher-quality than ICT's.
- **Large Language Models (LLMs):** We instruct LLMs to generate relevant queries for given passages. Since LLMs are strong in language understanding and generation, we consider the data quality to be better than both ICT and docT5query. We adopt ChatGLM3 [54] due to its impressive performance in various tasks. The prompt for query generation is shown in Appendix A.

For ICT and ChatGLM3, we generate a query for each positive document annotated by humans in the original datasets. For docT5query, we randomly sampled 500,000 passages from the corpus for query generation, since it is originally trained on the human annotated passages. In this way, we can align the training passages with human annotations and other annotations. Also, it's important to note that, despite employing different data generation methods,

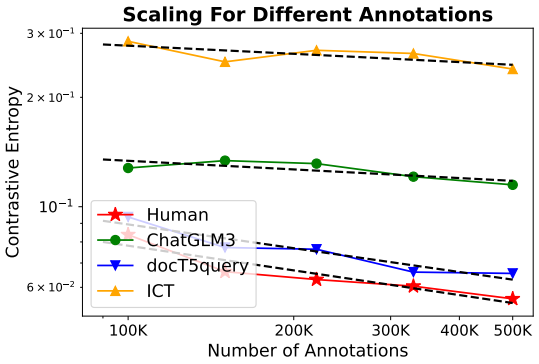

Figure 5: Scaling effects of annotation quality for retrieval performance on MS MARCO. Dashed lines are fitted using Eq. (7), which demonstrate the power-law scaling across different annotation methods. ChatGLM3 annotations exhibit the steepest slope and surpass human annotations at 500k.

our evaluations consistently utilize the human-annotated development set. The results are reported in Figure 5.

We can see that the retrieval performance scales with respect to different annotation qualities. Comparing the three methods of query generation, the log-linear curve of ICT exhibits the smallest slope. This observation aligns with our expectation that ICT is a weak supervision method and limits the enhancements for retrieval models when we increase the data size. The data quality from ChatGLM3 is better, but not as good as docT5query. This is because that docT5query has been finetuned on this dataset while ChatGLM3 is used in a zero-shot manner. Moreover, we use a 6B ChatGLM3 instead of a very large model, which may also result in its sub-optimal performance. Among all these methods, human annotations lead to the best-performing models. Therefore, there is still a large room for improvment about using large language models to generate pseudo training data.

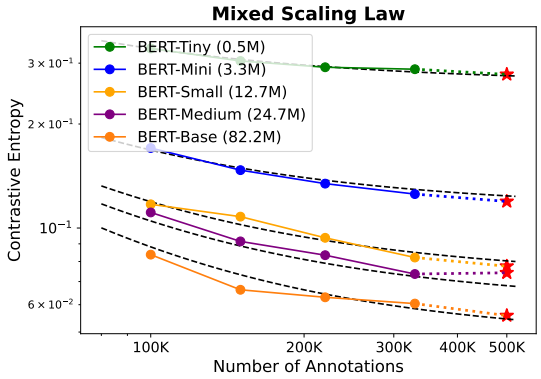

Figure 6: Modeling the joint effects of model size and data size on retrieval performance using a unified scaling law. Solid dots are used for fitting, while the red stars are performance to predict. The dashed lines are fitted with Eq. (8) and closely aligns with the observed data.

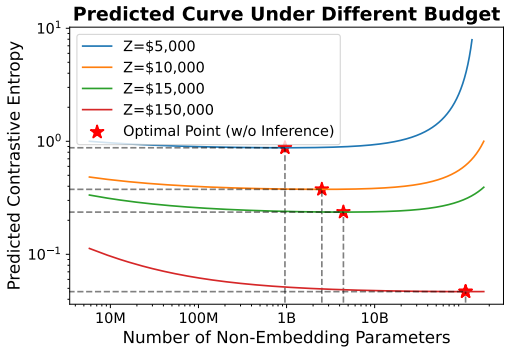

Figure 7: Predicted contrastive entropy for different model sizes under varying cost budgets, excluding inference costs. With an increase in model size, performance initially improves due to higher data efficiency of larger models, but eventually degenerates because of limited data annotation.

4.4 Model-Data Joint Laws

We combine the above observations into a single function that can characterize the joint effects of model size and data size. Inspired by the scaling laws of LLMs [22], we employ the following equation to describe the scaling effect:

$$L(N, D) = \left[\left(\frac{A}{N}\right)^{\frac{\alpha}{\beta}} + \frac{B}{D}\right]^{\beta} + \delta \tag{8}$$

$$A \approx 3.6 \times 10^4,\ B \approx 7.1 \times 10^3 \tag{9}$$

$$\alpha \approx 0.56,\ \beta \approx 1.31,\ \delta \approx 0.03 \tag{10}$$

where N, D represents the model size and data size, respectively, and $A, B, \alpha, \beta, \delta$ are coefficients. We employ results with different model sizes and data sizes to estimate the coefficients. Figure 6 illustrates the actual contrastive entropy and the predictions. In this figure, solid dots represent the data used for curve fitting, while the dashed line indicates the resulting fitted curve. The red stars denote data points utilized to evaluate the accuracy of our predictions. We can see that the predictions relatively are close to the real values.

5 APPLICATION IN BUDGET ALLOCATION

In this section, we showcase a potential application of the scaling laws for dense retrieval observed in our experiments. We use Eq. 8 in this section.

We attempt to estimate the comprehensive cost associated with the lifecycle of dense retrieval models, including data annotation, model training, and model inference. The total cost of training a model with N parameters using D data points is given by:

$$Z(N, D) = Z_{\text{data}} \cdot D + Z_{\text{train}} \cdot N + Z_{\text{infer}} \cdot N \tag{11}$$

Here, $Z_{\text{data}}, Z_{\text{train}}, Z_{\text{infer}}$ represent cost factors corresponding to annotations, training, and inference, respectively.

Now we estimate the approximate values for $Z_{\text{data}}, Z_{\text{train}}, Z_{\text{infer}}$. The cost of human annotations is approximated at \$0.6 per query-passage pair [3]. For computational costs, according to previous studies [7, 22], the training and inference computation for Transformer can be assumed by $6N$ and $2N$ FLOPs, respectively. We refer to common cloud computing and the price for using an A100 80G GPU is assumed to be \$3.93 per hour[3], with the peak computational power around 312 TFLOPs. For the training phase, we assume that the model is trained for 10,000 steps on a single A100 GPU. At each step, the model encodes a query, a positive passage and a negative passage with a batch size of 256. Each query is around 30 tokens and each passage is around 60 tokens. For the inference phase, we assume that the model is employed in a web search engine. Based on public statistics, we assume that there are around 30 trillion web pages in Google's index[4]. The inference cost for a dense retrieval model predominantly involves encoding the entire corpus. We estimate that each web page contains approximately 512 tokens. We assume the GPU utilization efficiency is 25%, then we have

$$Z_{\text{data}} \approx 0.6 \tag{12}$$

$$Z_{\text{train}} \approx \frac{10000 \times (30 + 2 \times 60) \times 256 \times 6 \times 3.93}{312T \times 3600 \times 25\%} = 3.22 \times 10^{-8} \tag{13}$$

$$Z_{\text{infer}} \approx \frac{30 \times 10^{12} \times 512 \times 2 \times 3.93}{312T * 3600 \times 25\%} = 0.43 \tag{14}$$

We first excludes the cost of inference and only focuses on annotation and training. Figure 7 shows the predicted contrastive entropy against model size under different cost budget. It is clear that for a fixed cost budget, as the training model size increases, the predicted retrieval performance initially improves and then slowly degenerates. The improvement is because that a relatively larger model is more data-efficient and can exhibit stronger ranking performance than smaller models. Nevertheless, when models are too large, only limited budget can be used for data annotation. The data limitations make the performance degenerate. Overall, if we do not consider inference cost, for a 20,000\$ budget, it is optimal to train a model with 13 billion parameters. This is primarily due to that

[3]https://cloud.google.com/compute/gpus-pricing
[4]From https://en.wikipedia.org/wiki/Google_Search, the estimated size of Google's index is around 30 trillion in 2012.

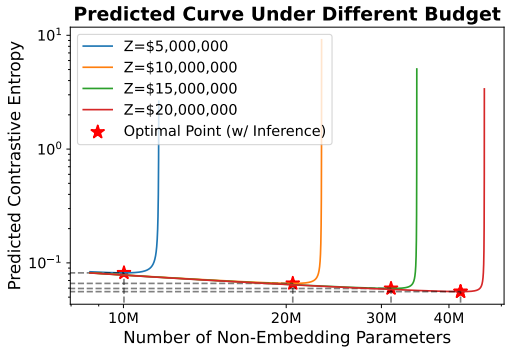

Figure 8: Predicted contrastive entropy for different model sizes under varying cost budgets, including inference costs. Inference is much more costly than training and results in small models be the optimum.

larger models are more data-efficient and that human annotation is significantly more expensive than training. Therefore, under a limited budget, maximizing model size can yield better results.

We further include the inference costs into this analysis. The result is shown in Figure 8. It is clear that the optimal model size significantly decreases to only million-scale parameters, even under a larger budget. This is because that the inference cost is huge compared to training cost ($Z_{\text{infer}} \gg Z_{\text{train}}$) and that the small models are more inference-efficient. A billion-scale model will make the inference cost prohibitively high.

6 LIMITATION AND FUTURE WORK

This study pioneers the investigation of scaling laws in dense retrieval. We cover major factors like model scale, datasets, training volume, and annotation methods. Several other aspects remain unexplored and can be addressed in future research.

Our experiments utilize contrastive entropy as the evaluation metric due to its continuity, which addresses the discrete nature of ranking metrics and facilitates the derivation of scaling laws. Although we demonstrate a positive correlation between contrastive entropy and ranking performance, it is important to note that they are not equivalent. For instance, as shown in Figure 2, similar contrastive entropy scores do not guarantee similar ranking performance. Future research may explore alternative metrics that might offer a more direct correlation with ranking outcomes.

The training process in this work is based on random negative sampling and contrastive learning. We do not cover more sophisticated training techniques, such as hard negative sampling [53, 56], distillation [19], and contrastive pre-training [13, 14, 20]. These methods could potentially influence the scaling behaviors observed and should be investigated in the future.

We focus on a common dense retrieval architecture where text is mapped to a single dense vector of fixed dimensionality. However, some researchers have experimented with variations in this architecture, such as mapping to vectors of varying dimensions [42], multiple vectors [24, 28], or even sparse vectors [12, 32]. Future work could explore how these architectural modifications impact the scaling laws for dense retrieval.

Our evaluations are conducted within in-domain datasets. Although we also attempt out-of-domain test (not reported in the paper), the available datasets are relatively small and yield unstable results, making it challenging to draw robust conclusions. Thus, our results do not currently account for out-of-domain scenarios, and more extensive evaluation could be beneficial in future work.

While we try to assess scaling across various scales, our resources limit the maximum size of our models and the extent of data size. Future work could further evaluate scaling laws on an even larger scale with more extensive models and annotations.

7 CONCLUSION

This paper systematically investigates the scaling laws of dense retrieval. We conduct experiments on both Chinese and English datasets to assess the impact of model size, data size, and annotation methods on retrieval performance. By utilizing contrastive entropy as the metric, we observe a power law relationship between performance and both model size and data size across different annotation methods and datasets. We also show that the scaling laws help optimize training processes. For instance, the scaling laws is important to budget allocation management, as demonstrated in our experiments. Moreover, scaling laws allows to evaluate the efficacy of different annotation methods. As shown in our experiments, there is still a large improvement room for using large language models to generate relevance annotations. We believe scaling laws offer a systematic approach to assess and improve the training processes of ranking models. While this study has laid a foundation for future exploration in this area, further research is needed to expand our understanding of scaling laws across more varied domains, scales, architectures, and evaluations.

A APPENDIX

We use the following prompt for ChatGLM3 to generate queries for one passage. Note that {} is the placeholder for the actual passage.

```
Please generate 5 relevant queries according to the
    ↪ given passage for search purpose.
1. Each query should be relevant to the passage.
2. Each query should be around 10 to 20 words.
3. Please generate diverse queries.
4. Output in JSON format, with keys: "query1", "query2
    ↪ ", "query3", "query4", "query5".
5. Please respond in English. DO NOT use Chinese.
Passage: {}
```

Listing 1: ChatGLM3 Prompt for Query Generation.

ACKNOWLEDGMENTS

This work is supported by Quan Cheng Laboratory (Grant No. QCLZD202301).

REFERENCES

[1] 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* (2023).

[2] Lada A Adamic and Bernardo A Huberman. 2002. Zipf's law and the Internet. *Glottometrics* 3, 1 (2002), 143–150.

[3] Sophia Althammer, Guido Zuccon, Sebastian Hofstätter, Suzan Verberne, and Allan Hanbury. 2023. Annotating Data for Fine-Tuning a Neural Ranker? Current Active Learning Strategies are not Better than Random Selection. In *Proceedings*

of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region. 139–149.

[4] Newsha Ardalani, Carole-Jean Wu, Zeliang Chen, Bhargav Bhushanam, and Adnan Aziz. 2022. Understanding Scaling Laws for Recommendation Models. *arXiv preprint arXiv:2208.08489* (2022).

[5] Yinqiong Cai, Yixing Fan, Jiafeng Guo, Fei Sun, Ruqing Zhang, and Xueqi Cheng. 2021. Semantic Models for the First-Stage Retrieval: A Comprehensive Review. *ACM Transactions on Information Systems (TOIS)* 40 (2021), 1 – 42. https://api.semanticscholar.org/CorpusID:232147859

[6] Aidan Clark, Diego De Las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, et al. 2022. Unified scaling laws for routed language models. In *International Conference on Machine Learning*. PMLR, 4057–4086.

[7] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *ICLR*. https://openreview.net/pdf?id=r1xMH1BtvB

[8] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. 2023. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*. PMLR, 7480–7512.

[9] 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, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[10] Zhengxiao Du, Aohan Zeng, Yuxiao Dong, and Jie Tang. 2024. Understanding emergent abilities of language models from the loss perspective. *arXiv preprint arXiv:2403.15796* (2024).

[11] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086* (2021).

[12] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2288–2292.

[13] Luyu Gao and Jamie Callan. 2021. Condenser: a Pre-training Architecture for Dense Retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 981–993. https://doi.org/10.18653/v1/2021.emnlp-main.75

[14] Luyu Gao and Jamie Callan. 2022. Unsupervised corpus aware language model pre-training for dense passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 2843–2853. https://doi.org/10.18653/v1/2022.acl-long.203

[15] Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*. PMLR, 10835–10866.

[16] Alexander Gelbukh and Grigori Sidorov. 2001. Zipf and Heaps laws' coefficients depend on language. In *Computational Linguistics and Intelligent Text Processing: Second International Conference, CICLing 2001 Mexico City, Mexico, February 18–24, 2001 Proceedings 2*. Springer, 332–335.

[17] 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 arXiv:1712.00409* (2017).

[18] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=iBBcRUlOAPR

[19] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 113–122.

[20] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. *Transactions on Machine Learning Research* (2022). https://openreview.net/forum?id=jKN1pXi7b0

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*. PMLR, 4904–4916.

[22] 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).

[23] 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. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

[24] 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*. 39–48.

[25] Jonathan C Lansey and Bruce Bukiet. 2009. Internet Search Result Probabilities: Heaps' Law and Word Associativity. *Journal of Quantitative Linguistics* 16, 1 (2009), 40–66.

[26] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021. Pretrained transformers for text ranking: Bert and beyond. *Synthesis Lectures on Human Language Technologies* 14, 4 (2021), 1–325.

[27] Linyuan Lü, Zi-Ke Zhang, and Tao Zhou. 2010. Zipf's law leads to Heaps' law: Analyzing their relation in finite-size systems. *PloS one* 5, 12 (2010), e14139.

[28] 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 (2021), 329–345.

[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 *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 1075–1088. https://doi.org/10.18653/v1/2021.eacl-main.92

[30] Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, and Xueqi Cheng. 2022. Pre-Train a Discriminative Text Encoder for Dense Retrieval via Contrastive Span Prediction. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 848–858. https://doi.org/10.1145/3477495.3531772

[31] Yixiao Ma, Yueyue Wu, Qingyao Ai, Yiqun Liu, Yunqiu Shao, Min Zhang, and Shaoping Ma. 2023. Incorporating Structural Information into Legal Case Retrieval. *ACM Transactions on Information Systems* 42, 2 (2023), 1–28.

[32] Antonio Mallia, Omar Khattab, Torsten Suel, and Nicola Tonellotto. 2021. Learning passage impacts for inverted indexes. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1723–1727.

[33] Mark EJ Newman. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary physics* 46, 5 (2005), 323–351.

[34] 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*.

[35] Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y Zhao, Yi Luan, Keith B Hall, Ming-Wei Chang, et al. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 9844–9855. https://doi.org/10.18653/v1/2022.emnlp-main.669

[36] Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019. From doc2query to docTTTTTquery. *Online preprint* 6, 2 (2019).

[37] Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375* (2019).

[38] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. 2023. Combined scaling for zero-shot transfer learning. *Neurocomputing* 555 (2023), 126658.

[39] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 5835–5847. https://doi.org/10.18653/v1/2021.naacl-main.466

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[41] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*. PMLR, 28492–28518.

[42] Nils Reimers and Iryna Gurevych. 2021. The Curse of Dense Low-Dimensional Information Retrieval for Large Index Sizes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint*

Conference on Natural Language Processing (Volume 2: Short Papers), Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 605–611. https://doi.org/10.18653/v1/2021.acl-short.77

[43] 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.

[44] Guilherme Moraes Rosa, Luiz Bonifacio, Vitor Jeronymo, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. No parameter left behind: How distillation and model size affect zero-shot retrieval. *arXiv preprint arXiv:2206.02873* (2022).

[45] 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*. Association for Computational Linguistics, Seattle, United States, 3715–3734. https://doi.org/10.18653/v1/2022.naacl-main.272

[46] Yunqiu Shao, Yueyue Wu, Yiqun Liu, Jiaxin Mao, and Shaoping Ma. 2023. Understanding Relevance Judgments in Legal Case Retrieval. *ACM Transactions on Information Systems* 41, 3 (2023), 1–32.

[47] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 3645–3650. https://doi.org/10.18653/v1/P19-1355

[48] 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. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1441–1451. https://doi.org/10.18653/v1/P19-1139

[49] Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022. Gpl: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 2345–2360. https://doi.org/10.18653/v1/2022.naacl-main.168

[50] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research* (2022). https://openreview.net/forum?id=yzkSU5zdwD Survey Certification.

[51] Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. RetroMAE: Pre-Training Retrieval-oriented Language Models Via Masked Auto-Encoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 538–548. https://doi.org/10.18653/v1/2022.emnlp-main.35

[52] Xiaohui Xie, Qian Dong, Bingning Wang, Feiyang Lv, Ting Yao, Weinan Gan, Zhijing Wu, Xiangsheng Li, Haitao Li, Yiqun Liu, and Jin Ma. 2023. T2Ranking: A Large-scale Chinese Benchmark for Passage Ranking. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (, Taipei, Taiwan,) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 2681–2690. https://doi.org/10.1145/3539618.3591874

[53] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *International Conference on Learning Representations*. https://openreview.net/forum?id=zeFrfgyZln

[54] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: An Open Bilingual Pre-trained Model. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=-Aw0rrrPUF

[55] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. 2022. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12104–12113.

[56] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1503–1512.

[57] Gaowei Zhang, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Scaling Law of Large Sequential Recommendation Models. *arXiv preprint arXiv:2311.11351* (2023).