RE-AdaptIR: Improving Information Retrieval through Reverse Engineered Adaptation

William Fleshman
will.fleshman@jhu.edu
Johns Hopkins University
Baltimore, Maryland, USA

Benjamin Van Durme
vandurme@jhu.edu
Johns Hopkins University
Baltimore, Maryland, USA

Abstract

Large language models (LLMs) fine-tuned for text-retrieval have demonstrated state-of-the-art results across several information retrieval (IR) benchmarks. However, supervised training for improving these models requires numerous labeled examples, which are generally unavailable or expensive to acquire. In this work, we explore the effectiveness of extending reverse engineered adaptation to the context of information retrieval (RE-AdaptIR). We use RE-AdaptIR to improve LLM-based IR models using only unlabeled data. We demonstrate improved performance in both training domains and in zero-shot domains where the models have seen no queries. We analyze performance changes in various fine-tuning scenarios and offer findings of immediate use to IR practitioners.

CCS Concepts

• **Information systems** → **Information retrieval**; **Language models**.

Keywords

Neural IR; LLM; adapter-tuning; LoRA; fine-tuning

ACM Reference Format:
William Fleshman and Benjamin Van Durme. 2025. RE-AdaptIR: Improving Information Retrieval through Reverse Engineered Adaptation. 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, 5 pages. https://doi.org/10.1145/3726302.3730240

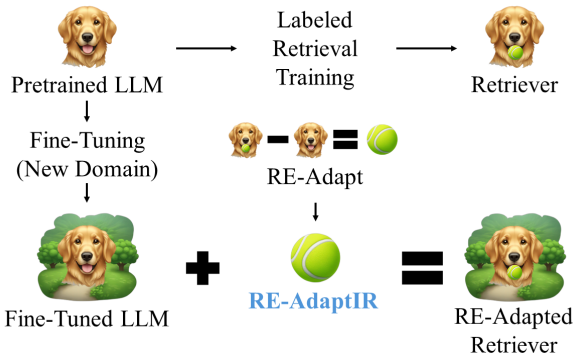

Figure 1: In RE-AdaptIR, RE-Adapt is extended to an existing retrieval model to isolate what was learned during labeled contrastive training. The pretrained model is fine-tuned on unlabeled in-domain documents and *readapted* for text retrieval. The new retriever outperforms the original on both in-domain and zero-shot information retrieval tasks.

1 Introduction

Information retrieval (IR) is a fundamental component of various modern applications, powering search engines, recommender systems, and various data analytics pipelines. Recently, large language models (LLMs) have achieved state-of-the art results on dense text retrieval, identifying and ranking the most relevant text for a given query by comparing learned vector representations of the text [9–12, 17, 23, 31]. The effectiveness of text retrieval have a direct impact on numerous domains, including healthcare, finance, and social media, where accurate and timely access to information is critical. Retrieval is also critical in the context of retrieval augmented generation (RAG), enabling LLMs access to external resources when constructing a response [15]. For these reasons, we seek a practical and efficient approach to improve retrieval models.

Supervised fine-tuning of LLMs for text retrieval tasks has become a widely adopted approach, leveraging their pretrained language understanding capabilities to achieve state-of-the-art results on various benchmarks [10, 17, 31]. However, adapting an LLM for text retrieval requires labeled datasets, with numerous example queries and documents both related and unrelated to forming a helpful response. This poses a significant challenge to improving these systems, as data annotation or synthetic generation can be too expensive, difficult, and error-prone [3, 5]. Making matters worse, fine-tuning an existing LLM on new domains can cause *forgetting*, a decreased performance on previously capable tasks [13, 19]. Fleshman and Van Durme [4] recently proposed reverse engineered adaptation (RE-Adapt), an approach for solving similar dilemmas faced when fine-tuning existing instruction-tuned models. Here we introduce RE-AdaptIR, an extension of RE-Adapt to IR models, which leverages the available unlabeled data to improve existing text-retrieval LLMs (Figure 1). Specifically we:

- Extend RE-Adapt to the information retrieval setting and apply RE-AdaptIR to two state-of-the-art text retrieval models: RepLLaMA and e5-Mistral;
- Demonstrate improved performance both in-domain and zero-shot across 14 datasets; and
- Explore the importance of fine-tuning on data relevant to test-time queries and the impact different scenarios have on performance.

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730240

2 Background

2.1 Retrieval Models

The transformer architecture is a natural choice for text retrieval models, as it embeds text into dense vector representations [28]. LLMs pretrained on massive amounts of text have demonstrated the ability to capture useful semantic meaning in these vector representations [10, 22, 26]. To ensure these models are capable for retrieval, a secondary fine-tuning stage is used to encourage the model to map similar texts to similar vectors [10, 17, 31]. This is generally done via some form of supervised contrastive training such as with InfoNCE [27]. Given a query representation Q, document representations $\{D_N\}$, and a similarity function ϕ, models are optimized by minimizing the InfoNCE loss:

$$\mathcal{L} = -\log \frac{\exp(\phi(Q, D^+))}{\exp(\phi(Q, D^+)) + \Sigma_{D_i^- \in \{D_N\}} \exp(\phi(Q, D_i^-))}, \quad (1)$$

where D^+ is a document relevant to Q and D^- an irrelevant document. After training, the models can be used to create a database of document vectors, and rank them given a query representation using the similarity function ϕ. Related documents should have higher similarity than those unrelated. We experiment with two such models in this work: RepLLaMA [17] and e5-Mistral [29].

RepLLaMA. Ma et al. [17] introduced RepLLaMA[1] to demonstrate that state-of-the-art LLMs could surpass the previous results achieved with smaller retrieval models, especially when evaluated zero-shot on datasets not seen during training. They construct RepLLaMA by fine-tuning LLaMA-2-7B [26] on approximately 500k labeled examples from the training split of MS-MARCO [1].

e5-Mistral. In a similar fashion, Wang et al. [29] fine-tune the Mistral-7B LLM [10] with a focus on synthetic data. They also trained with a combination of MS-MARCO, and multiple other labeled datasets. The resulting e5-Mistral[2] model achieved state-of-the-art results on several text-based retrieval benchmarks [29].

Equation 1 works well for optimizing retrieval models, but the requirement to have query-document pairs is highly constraining. For both models, labeled data was needed to achieve the best results, and it is unclear how to effectively incorporate the copious amount of unlabeled text additionally available. For example, MS-MARCO contains almost 9 million passages, but only a fraction of this data is used in training, due to the limited number of associated queries available. In this work, we use RE-AdaptIR to leverage this unlabeled data to improve the performance of these systems.

2.2 Reverse Engineered Adaptation

Fleshman and Van Durme [4] introduced reverse engineered adaptation (RE-Adapt) as a new method to efficiently update instruction-tuned models with unlabeled data lacking the previously required instruction-tuning annotations [20, 21, 30]. RE-Adapt isolates what has been learned from instruction-tuning by taking the difference between the weights of the instruction-tuned and pretrained versions of a model. This difference can be thought of as an adapter [6] or as a multi-task version of task-vectors [8]. Given this RE-Adapter Δ, the pretrained weights Θ can be fine-tuned with a new *knowledge adapter* Ψ without impacting the previous instruction-tuning. Finally, the model can be re-instantiated with weights $\Theta + \alpha\Psi + \beta\Delta$ where α and β are *partial adaptation* scalars used to control the strength of fine-tuning [4]. The authors show that RE-Adapt improves the performance of instruction-tuned models in the new domain while preserving or improving performance out-of-domain.

3 Re-AdaptIR

In this work, we explore the effectiveness of RE-Adapt for text-based information retrieval. In RE-Adapt, instruction-tuned models still leverage the pretraining capabilities of next-token prediction, but most text retrieval models do not. RepLLaMA and e5-Mistral both discard the next-token predictor from their respective LLMs and fine-tune the model to produce a single vector representation per document [17, 29]. It is therefore unclear whether continued fine-tuning of the pretrained LLM with next-token prediction will improve the existing down-stream retrieval performance.

To answer this question, we first use LoRA [7] to fine-tune the weights Θ of a pretrained LLM on unlabeled documents from a new domain, producing adapter weights Ψ. This process attempts to learn Ψ which minimizes the next-token prediction loss of the model using weights $\Theta + \Psi$ [3]. Next, we consider a retrieval model with weights Λ, fine-tuned from the same pre-trained model to minimize Equation 1, without the next-token-predictor weights. We construct a RE-AdaptIR Δ by finding the difference between the retrieval and pretrained weights: $\Delta = \Lambda - \Theta$, isolating the changes derived from retrieval training. We then construct a RE-Adapted IR model with weights Ω using a layer-wise linear combination of the pretrained weights, the adapter weights, and the RE-AdaptIR weights:

$$\Omega = \Theta + 0.5\Psi + \Delta, \quad (2)$$

where the adapter Ψ is scaled by 0.5 to reduce interference [4]. Unlike, Fleshman and Van Durme [4], we do not scale Δ, resulting in an equivalent representation of Equation 2 in terms of the original retrieval model weights and the adapter capturing knowledge from the new data:

$$\Omega = \Lambda + 0.5\Psi. \quad (3)$$

4 Experiments

We first replicate the shared text retrieval experiments conducted by Ma et al. [17] and Wang et al. [29] and compare the base model performance before and after applying RE-AdaptIR. Retrieval is performed by finding the top-k most similar document vectors for each query vector. We then analyze how different fine-tuning scenarios impact the IR performance.

4.1 Adapter Details

We use parameter efficient fine-tuning [18] with LoRA [7] to adapt the pretrained LLaMA-2-7B [26] and Mistral-7B [10] LLMs to each dataset being evaluated. We use rank 32 adapters with a LoRA $\alpha = 64$, dropout of 0.05, and DoRA [16] enabled. Adapters are added to all attention key, query, and value layers as well as the

[1]The RepLLaMA model license can be found at https://huggingface.co/castorini/repllama-v1-7b-lora-passage.

[2]The e5-Mistral model license can be found at https://huggingface.co/intfloat/e5-mistral-7b-instruct.

[3]This sum is performed layer-wise. Layers without corresponding adapter weights are used without modification.

linear up and down projection layers for each model. We train each adapter for a single epoch on all passages from the dataset under evaluation using a learning rate of 2e-4 with the AdamW optimizer and linear scheduling. A single Nvidia A100 GPU was used for all training and experimentation. As in Fleshman and Van Durme [4], we use a scalar of 0.5 with our knowledge adapters to minimize interference between the adapters and existing retrieval ability.

4.2 Datasets

Both RepLLaMA and e5-Mistral utilized the MS-MARCO [1] dataset as part of their training data, and we include it in our evaluations to help measure any benefits from using RE-AdaptIR in-domain. Specifically, RE-AdaptIR allows for fine-tuning over the entire 8.84M passages, where only a subset of those passages was used for retrieval training due to the limited availability of query-passage pairs [17, 29]. Additionally, we use the same 13 public datasets from the BeIR IR benchmark [24] used by Ma et al. [17] to assess RepLLaMA's zero-shot performance across a diverse set of IR tasks. Of these, we note that the training splits of FEVER [25], HotPotQA [32], NQ [14], and Quora [2] were also used by Wang et al. [29] in the training of e5-Mistral, providing more in-domain insight to our experiments. These datasets are zero-shot for RepLLaMA, as are the remaining BeIR datasets for both models. We use the same prompts used by Ma et al. [17] and Wang et al. [29] for each dataset.[4]

4.3 Main Result

Our main results are compiled in Table 1. We see that RE-AdaptIR improves performance in the majority of cases, increasing the average zero-shot nDCG@10 by 1.1% (0.6 points) and 4.1% (1.9 points) for RepLLaMA and e5-Mistral respectively. Individual dataset performance varies, with increases of up to 7 points. **Importantly, these significant[5] performance gains required no additional labeled data** and are achieved by simply fine-tuning the pretrained model over the document database being used for retrieval. Using an approach like LoRA allows for efficiently gaining these improvements on commodity hardware [7]. The default partial adaptation scalar of 0.5 was used for this experiment, but we note that optimizing this value per dataset does improve results. While not applicable to our zero-shot analysis, performance can be further boosted by using withheld queries to tune all hyperparameters for LoRA training.

We notice the few cases where performance was reduced tend to occur with the larger corpora. We plot this relationship in Figure 2 and observe a slightly negative correlation. This relationship is purely observational and likely caused by latent topic or task diversity among the larger datasets used in these experiments. A larger sample size would be necessary to confirm if this relationship holds in general, but we include the observation for consideration.

4.4 Performance Analysis

We have demonstrated that RE-AdaptIR improves the performance of both retrieval models by simply fine-tuning the pretrained model on additional unlabeled data. Now, we conduct further experimentation to answer additional questions of importance to practitioners.

Dataset	RepLLaMA Base	RepLLaMA RA	e5-Mistral Base	e5-Mistral RA
MS-MARCO	**46.5**	46.1	36.5	**40.1**
FEVER	**84.0**	83.8	85.1	**87.7**
HotPotQA	67.2	**67.6**	72.5	**73.4**
NQ	61.8	**62.1**	**53.3**	52.4
Quora	80.1	**82.8**	85.6	**88.0**
Arguana	52.3	**52.6**	52.0	**59.0**
Climate-FEVER	**30.8**	30.4	24.9	**31.4**
DBPedia	43.4	**43.5**	**47.2**	47.1
FiQA	44.2	**45.5**	49.9	**52.3**
NFCorpus	38.0	**38.6**	39.6	**40.8**
SCIDOCS	17.7	**18.3**	18.6	**18.7**
SciFact	74.5	**76.3**	71.4	**73.3**
TREC-COVID	84.0	**85.6**	**83.9**	81.0
Touche-2020	**27.5**	27.0	29.0	**30.1**
Average	53.7	**54.3**	53.5	**55.4**
Average Z-Shot	54.3	**54.9**	46.3	**48.2**

Table 1: nDCG@10 across test splits for MS-MARCO and BeIR datasets. Results highlighted in orange indicate the dataset's train split was used for training the corresponding model and are not zero-shot. *Base* is the unmodified model and *RA* is the model RE-Adapted after fine-tuning on the domain.

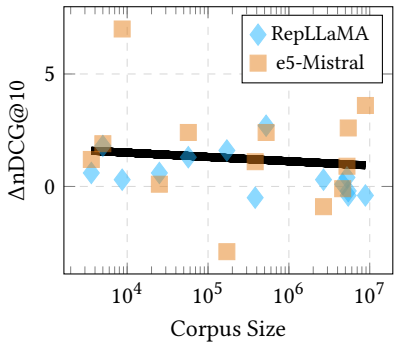

Figure 2: The observed relationship between the corpus size and the change in retrieval performance after fine-tuning.

Are queried passages all that matter? One reason the in-domain results could be better than baseline is because the passages being queried for at test-time are included in the fine-tuning data, although their corresponding queries are not. We test this hypothesis by training two additional knowledge adapters, one which sees no test-time passages, and one that sees only test-time passages. We compare the evaluation results with the original adapter fine-tuned on all the passages (Table 2). We find that neither of the subsets is always best, indicating that having in-domain data is more important than specifically fine-tuning on the passages being queried for at test-time. Retreival systems generally have access to the documents being queried, so fine-tuning on these passages is practical. However, this result indicates that developers of IR models can improve

[4]The datasets and licenses used in this work can be found at https://github.com/beir-cellar/beir.
[5]Significance tests result in p-values of 0.038 (RepLLaMA) and 0.002 (e5-Mistral).

RE-Adapted Model	Arguana			FiQA			NFCorpus			SciFact		
	w/o	w/	both	w/o	w/	both	w/o	w/	both	w/o	w/	both
RepLLaMA	+0.5	+0.2	+0.3	+1.0	+1.2	+1.3	+0.3	+0.5	+0.6	+0.1	+0.5	+1.8
e5-Mistral	+3.9	+5.1	+7.0	+2.0	+2.8	+2.4	+0.6	+1.1	+1.2	+2.3	+1.0	+1.9

Table 2: The change in nDCG@10 from the original retriever when the pretrained model is fine-tuned using only documents with (w/) or without (w/o) corresponding queries in the test set, or with (both) subsets. Performance improves in all cases.

the model on specific domains before deployment, without access to the specific documents relevant to their end users.

Does any unlabeled data work? Next, we explore the importance of using domain-specific data for fine-tuning. We established that fine-tuning data can be a different set than the data being queried over, but can it be any set? We repeat our main experiment across the BeIR datasets, but using only the knowledge adapter trained on MS-MARCO, our largest corpus. We compare the resulting performance with the original RepLLaMA and e5-Mistral baselines as well as the RE-Adapted models fine-tuned on the domain specific data (Table 3). We observe that additional fine-tuning with MS-MARCO improves RepLLaMA by an average of 0.2 points but is still 0.5 points below the average performance when using in-domain data. For e5-Mistral however, we see that the MS-MARCO fine-tuning results in a significant increase of 2 points over baseline on average, 0.3 points above what is achieved with in-domain data. The larger gain with e5-Mistral is likely due to RepLLaMA's use of MS-MARCO as the majority of its original training data, while e5-Mistral only used a subset to supplement the otherwise synthetically generated data [17, 29]. In both cases, the extra data improved the original performance and indicates that retrieval models can generally benefit from additional unlabeled training using RE-AdaptIR.

Dataset	RepLLaMA		e5-Mistral	
	Base	Domn	Base	Domn
FEVER	+0.1	+0.3	-0.1	-2.7
HotPotQA	-0.4	-0.8	+0.5	-0.4
NQ	-0.1	-0.4	+7.0	+7.9
Quora	+2.0	-0.7	+2.4	0.0
Arguana	+0.8	+0.5	+10.1	+3.1
Climate-FEVER	+1.0	+1.4	-1.1	-7.6
DBPedia	+0.1	0.0	+0.5	+0.6
FiQA	+0.2	-1.1	+1.7	-0.7
NFCorpus	0.0	-0.6	+0.7	-0.5
SCIDOCS	+0.1	-0.5	+0.6	+0.5
SciFact	-0.9	-2.7	+2.1	+0.2
TREC-COVID	-0.4	-2.0	+2.6	+5.5
Touche-2020	-0.1	+0.4	-0.5	-1.6
Average	+0.2	-0.5	+2.0	+0.3

Table 3: Change in nDCG@10 when RE-AdaptIR is applied with pretrained model fine-tuned on MS-MARCO instead of the domain under evaluation. *Base* indicates the change with respect to the original model, *Domn* the change with respect to the model RE-Adapted on the evaluated domain.

Does additional training improve results? Our largest performance decrease was e5-Mistral RE-Adapted with one epoch of TREC-COVID data. We repeat the experiment by continuing to train and evaluate performance after each additional epoch of fine-tuning (Figure 3). RE-AdaptIR continues to improve, surpassing e5-Mistral after 4 epochs. This result is further indication that tuning hyperparameters with held-out queries will boost the performance of RE-AdaptIR. Although our main focus and results demonstrate the approach also works well using only unlabeled documents.

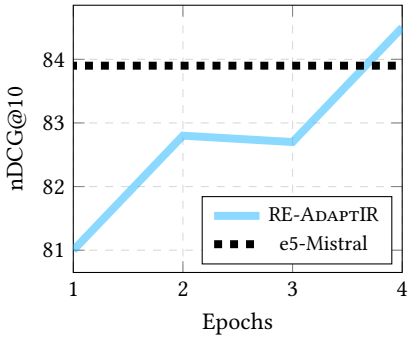

Figure 3: Performance on TREC-COVID as e5-Mistral continues fine-tuning on unlabeled data through RE-AdaptIR. The baseline performance is surpassed after 4 epochs of training.

5 Conclusion

In this work, we introduced RE-AdaptIR, an extension of RE-Adapt for using unlabeled data to improve the zero-shot and in-domain performance of text retrieval models. We demonstrated RE-AdaptIR improves two state-of-the-art models: RepLLaMA and e5-Mistral, using parameter-efficient fine-tuning over raw documents without access to related queries. We found that fine-tuning on the documents being queried for at test-time is not required, and we still see increased performance when RE-AdaptIR is used with unrelated documents from the same domain. RE-AdaptIR also improved baseline retrieval performance when fine-tuning was performed using additional unlabeled data from the models original training corpus, even though this data was out-of-domain with respect to the data being queried. Our main experiments assumed no access to labeled document-query pairs, but we showed that RE-AdaptIR can be further improved through the tuning of partial adaptation scalars or training epochs. Combined, these results enforce the wide applicability of our approach, and our findings ensure RE-AdaptIR is of immediate use to text-retrieval practitioners.

References

[1] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. arXiv:1611.09268

[2] DataCanary, hilfialkaff, Lili Jiang, Meg Risdal, Nikhil Dandekar, and tomtung. 2017. Quora Question Pairs. https://kaggle.com/competitions/quora-question-pairs

[3] Michael Desmond, Evelyn Duesterwald, Kristina Brimijoin, Michelle Brachman, and Qian Pan. 2021. Semi-Automated Data Labeling. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track (Proceedings of Machine Learning Research, Vol. 133)*, Hugo Jair Escalante and Katja Hofmann (Eds.). PMLR, 156–169.

[4] William Fleshman and Benjamin Van Durme. 2024. RE-Adapt: Reverse Engineered Adaptation of Large Language Models. arXiv:2405.15007

[5] Teodor Fredriksson, David Issa Mattos, Jan Bosch, and Helena Holmström Olsson. 2020. Data Labeling: An Empirical Investigation into Industrial Challenges and Mitigation Strategies. In *Product-Focused Software Process Improvement*, Maurizio Morisio, Marco Torchiano, and Andreas Jedlitschka (Eds.). Springer Int'l Publishing, Cham, 202–216.

[6] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th Int'l Conf. on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 2790–2799.

[7] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Int'l Conf. on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

[8] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh Int'l Conf. on Learning Representations*.

[9] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. arXiv:2112.09118

[10] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825

[11] 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 *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 6769–6781. doi:10.18653/v1/2020.emnlp-main.550

[12] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. ACM, New York, NY, USA, 39–48. doi:10.1145/3397271.3401075

[13] Suhas Kotha, Jacob Springer, and Aditi Raghunathan. 2024. Understanding Catastrophic Forgetting in Language Models via Implicit Inference. In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*. https://openreview.net/forum?id=wkQy8mLIb9

[14] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (2019), 452–466. doi:10.1162/tacl_a_00276

[15] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th Int'l Conf. on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 793, 16 pages.

[16] Shih-yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. DoRA: Weight-Decomposed Low-Rank Adaptation. In *Forty-first Int'l Conf. on Machine Learning*.

[17] Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-Tuning LLaMA for Multi-Stage Text Retrieval. In *Proceedings of the 47th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. ACM, New York, NY, USA, 2421–2425. doi:10.1145/3626772.3657951

[18] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods. https://github.com/huggingface/peft.

[19] Michael McCloskey and Neal J. Cohen. 1989. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. Psychology of Learning and Motivation, Vol. 24. Academic Press, 109–165. doi:10.1016/S0079-7421(08)60536-8

[20] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-Task Generalization via Natural Language Crowdsourcing Instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3470–3487. doi:10.18653/v1/2022.acl-long.244

[21] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv* abs/2203.02155 (2022).

[22] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

[23] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3982–3992. doi:10.18653/v1/D19-1410

[24] 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 *Thirty-fifth Conf. on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[25] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 809–819. doi:10.18653/v1/N18-1074

[26] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288

[27] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. arXiv:1706.03762

[29] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving Text Embeddings with Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 11897–11916. doi:10.18653/v1/2024.acl-long.642

[30] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned Language Models are Zero-Shot Learners. In *Int'l Conf. on Learning Representations*.

[31] Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. 2024. FollowIR: Evaluating and Teaching Information Retrieval Models to Follow Instructions. arXiv:2403.15246

[32] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 2369–2380. doi:10.18653/v1/D18-1259