Investigating Task Arithmetic for Zero-Shot Information Retrieval*

Marco Braga[1,2,*], Pranav Kasela[1], Alessandro Raganato[1] and Gabriella Pasi[1]

[1]Università degli Studi di Milano-Bicocca, Milano
[2]Politecnico di Torino, Dipartimento di Automatica e Informatica DAUIN, Corso Duca degli Abruzzi, Torino

Abstract

Large Language Models (LLMs) have shown impressive capabilities in zero-shot scenarios, i.e. where no task-specific labelled data is provided. However, their performance tends to degrade when applied to previously unseen domains or tasks, primarily due to differences in term distributions. In this paper, we explore Task Arithmetic, a simple yet effective method that allows knowledge transfer between domain-specific and retrieval models. By leveraging basic mathematical operations, such as vector addition and subtraction, we construct LLMs that incorporate domain-specific knowledge into existing re-ranking models, without the need for additional Fine-Tuning. Experimental results across eight benchmark datasets, including scientific, biomedical, and multilingual corpora, show that our method consistently enhances re-ranking performance, with gains up to 18% in NDCG@10 and 15% in P@10. We make our code publicly available at https://github.com/DetectiveMB/Task-Arithmetic-for-ZS-IR.

Keywords

Task Arithmetic, Domain-specific IR, Multilingual IR, Zero-Shot

1. Introduction

Large Language Models (LLMs) have shown state-of-the-art performance across a broad range of Natural Language Processing (NLP) tasks [2, 3], including Information Retrieval (IR) [4]. Their ability to capture rich semantic representations from large-scale, unlabeled corpora has enabled their applications to different IR tasks, such as document re-ranking [5, 6, 7, 8, 9], query expansion [10, 11], and the generation of synthetic data [12, 13]. Notably, LLMs excel in zero-shot scenarios [14], where models are evaluated on domains unseen during training, without requiring any domain-specific labelled data. Despite these advantages, domain mismatch remains a critical challenge due to differences in terminology, distributional properties, and task formulations [15]. The BEIR benchmark [16], which spans different tasks and domains, provides a heterogeneous framework for evaluating zero-shot IR performance. A common approach to the BEIR dataset involves fine-tuning a model on a large-scale IR dataset, such as MS MARCO [17], and applying it in a zero-shot setting to unseen domains [18]. While this strategy can yield competitive results [16], achieving robust generalization across all domains with a single model remains challenging [19]. This is largely due to the requirement for domain-specific datasets (and annotations) and the computational cost associated with retraining or adapting models for each new task or domain [20, 21, 22, 23, 24, 25]. In this work, we investigate Task Arithmetic [26] as an alternative strategy for Fine-Tuning LLMs to unseen tasks, domains and language adaptation in the context of IR. Task Arithmetic enables the transfer of domain-specific knowledge by combining model parameters through simple vector operations, such as addition and subtraction, without requiring any additional Fine-Tuning. Specifically, we define Task Vectors that represent the difference between a domain or task-specific model and its original pre-trained version. These vectors are then added to IR-tuned baselines to define a new model that integrates both retrieval capabilities and task-specific expertise. We evaluate our approach across eight publicly available datasets covering scientific, biomedical, and

IIR2025: 15th Italian Information Retrieval Workshop, 3th - 5th September 2025, Cagliari, Italy
*This is an extended abstract of [1] (accepted at SIGIR 2025)
*Corresponding author.
m.braga@campus.unimib.it (M. Braga); pranav.kasela@unimib.it (P. Kasela); alessandro.raganato@unimib.it (A. Raganato); gabriella.pasi@unimib.it (G. Pasi)
© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

CEUR-WS.org/Vol-4026/paper9.pdf

multilingual retrieval tasks. Our study includes six LLM architectures, ranging from encoder-only to encoder-decoder and decoder-only models, with parameter counts spanning from 66M to 7B. Results show that Task Arithmetic yields consistent improvements over strong IR baselines, with gains of up to 18% in NDCG@10 and 15% in P@10. These findings highlight the effectiveness of Task Arithmetic as a lightweight, training-free, and modular approach for zero-shot adaptation in IR, offering a practical alternative to conventional Fine-Tuning, especially when labeled data is scarce or unavailable.

2. Methodology

Adapting Large Language Models (LLMs) to specialized domains or tasks typically demands extensive Fine-Tuning, which requires high computational resources and data-related overhead. To mitigate these costs, particularly in scenarios characterized by frequent domain shifts or limited access to labeled data, recent works explore weight interpolation and model merging [27, 28]. These approaches exploit the observation that models fine-tuned on related tasks tend to occupy compatible regions of parameter space [29, 30], thereby enabling weight merging strategies such as averaging or interpolation to preserve or even enhance downstream performance [31, 32]. Among these methods, Task Arithmetic [26] has emerged as a particularly lightweight and effective approach. Unlike traditional adaptation techniques that require gradient-based updates [33, 34, 35, 36], Task Arithmetic enables model composition through simple mathematical operations, such as addition and subtraction, computing and combining difference parameter vectors between models. The Task Vector, obtained by subtracting the parameters of the pre-trained model from those of the domain- or task-specific variant, encapsulates the domain adaptation knowledge learned during Fine-Tuning and can be applied to another model by direct addition [37, 38]. Despite its promising effectiveness in various NLP tasks [26, 38], Task Arithmetic has not been investigated in the context of Zero-Shot Information Retrieval (IR). Most existing domain adaptation methods in IR rely on supervised fine-tuning [19, 39, 40], which still depend on annotated domain-specific data. In contrast, Task Arithmetic offers a plug-and-play solution: it leverages publicly available domain-adapted models and reuses them without any retraining, enabling efficient model transfer under zero-shot conditions. We formalize our application of Task Arithmetic to IR as follows. Let $\Theta_0 = \{(\theta_1)_0, \ldots, (\theta_N)_0\}$ denote the parameters of a pre-trained LLM. Fine-tuning this model on a generic IR task (e.g., on the MS-MARCO benchmark [17]) produces $\Theta_T = \{(\theta_1)_T, \ldots, (\theta_N)_T\}$, while fine-tuning Θ_0 on a specific domain yields $\Theta_D = \{(\theta_1)_D, \ldots, (\theta_N)_D\}$. We define the Task Vector τ_D for domain D as follows:

$$\tau_D = \{\tau_1, \ldots, \tau_N\}, \quad \text{where} \quad \tau_i = (\theta_i)_D - (\theta_i)_0.$$

This vector τ_D represents the domain-specific shift in the parameter space. To create a domain-aware IR model Θ', we add τ_D to the IR-tuned model Θ_T:

$$\Theta' = \{\theta'_i = (\theta_i)_T + \alpha\tau_i\}_{i=1}^{N}$$

The scaling factor $\alpha \in \mathbb{R}$ controls how much of the domain vector is injected. If $\alpha = 0$, then Θ' defaults to Θ_T. Setting $\alpha > 0$ adds the specialized knowledge, while $\alpha < 0$ subtracts it. In practice, α is a hyperparameter that can be tuned on a small development set.

3. Results and Discussion

In this Section, we describe the datasets, evaluation metrics, and models used to assess the effectiveness of our proposed approach. Then we present the results of our evaluation, which spans scientific, biomedical, and multilingual scenarios to show the potential of Task Arithmetic for zero-shot IR.

3.1. Datasets and Models

We assess zero-shot Information Retrieval performance on eight publicly available datasets. Four of these, i.e. TREC-COVID [41], NFCorpus [42], SCIDOCS [43], and SciFact [44], are drawn from

the BEIR benchmark [16] and target specialized domains and tasks such as biomedical retrieval and scientific citation prediction. Regarding language-specific retrieval, we additionally include GermanQuAD [45] and three multilingual datasets from the MIRACL benchmark [46], i.e. English, French, and Spanish. Retrieval effectiveness is measured using P@10, NDCG@10 and MAP@100. Statistical significance is determined via Bonferroni-adjusted, two-sided paired t-tests at a 99% confidence level. Significant gains over the best baseline are marked with an asterisk in all result tables. Our experiments span six LLM architectures covering a range of neural paradigms and models: `DistilBERT` [47] and `RoBERTa-base` [48] as encoder-only bi-encoders, `T5-base`, `T5-Large` [49], and `MT5-base` [50] as encoder-decoder cross-encoders, and `LLama-2-7b` [51] as a decoder-only LLM. For each pre-trained model, we compute Task Vectors by subtracting the weights of a publicly available domain- or language-specific fine-tuned model from those of its original pre-trained version. Specifically, we use `LLama2-MedTuned-7b` [52], `SciFive` [53], `BioMed-RoBERTa` [54], and `Bio-DistilBERT` [55] for the biomedical and scientific domains, as well as `MT5-base-german`, `MT5-base-spanish`, `MT5-base-french`, and `MT5-base-english` [56] for language adaptations. These Task Vectors are then added to models fine-tuned on MS-MARCO (`RankingGPT-Llama2-7b` [57], `MonoT5` [58], `msmarco-RoBERTa` [59], `msmarco-distilbert` [59], and `MT5-base-msmarco` [60]) [17]. To create the Task Arithmetic model Θ', in a fully zero-shot scenario, we set $\alpha = 1$. Furthermore, in a setting where a small set of labeled data is available, we tune the scaling factor α from 0.1 to 1.0 in steps of 0.1, selecting the optimal value based on the highest average retrieval performance over two development sets: the official NFCorpus development split and a 20% subset of SciFact training queries. Since GermanQuAD and MIRACL do not provide development sets, we apply a fully zero-shot scenario by not optimizing the value of α, i.e. $\alpha = 1$. All re-ranking experiments begin by retrieving the top 100 documents via BM25. The final rankings are then computed using a weighted sum of BM25 and LLM scores, with λ_{BM25} and λ_{LLM} optimized in $[0, 1]$ on the NFCorpus and SciFact development sets. We take the average score for all remaining datasets.

4. Results and Discussion

Model		SciFact			NFCorpus			SCIDOCS			TREC-COVID		
Re-ranker	Variant	P@10	NDCG@10	MAP@100	P@10	NDCG@10	MAP@100	P@10	NDCG@10	MAP@100	P@10	NDCG@10	MAP@100
BM25		.091	.691	.649	.247	.343	.154	.086	.165	.112	.764	.688	.085
`Llama-2-7B`	Θ_T: MS-MARCO (RankingLlama)	**.099**	**.770**	**.731**	**.265**	**.373**	**.170**	.096	.188	.129	.860	.810	.098
	Θ': Task Arithmetic ($\alpha = 1$)	.096	.757	.723	.265	.370	.167	.095	.185	.126	.858	.801	.098
	Θ': Task Arithmetic (optimized $\alpha = 0.8$)	.097	.765	.730	.262	.365	.165	.097	.189	.129	**.866**	**.812**	**.099***
`T5-Large`	Θ_T: MS-MARCO (Mono-T5)	**.095**	**.743**	**.709**	**.266***	**.368***	**.167**	.095	.182	.124	.784	.735	.092
	Θ': Task Arithmetic ($\alpha = 1$)	.092	.721	.692	.257	.356	.161	.096	.185	.124	.816	**.765**	.096
	Θ': Task Arithmetic ($\alpha = 0.9$)	.092	.727	.699	.259	.359	.162	**.098***	**.187***	**.126**	**.818**	.759	**.097***
`T5-base`	Θ_T: MS-MARCO (Mono-T5)	.096	.726	.684	.258	**.359**	**.162**	.090	.173	.118	.762	.712	.089
	Θ': Task Arithmetic ($\alpha = 1$)	.089	.686	.649	.250	.345	.156	.070	.136	.094	.764	.726	.088
	Θ': Task Arithmetic ($\alpha = 0.7$)	**.098**	**.748**	**.708***	**.259**	.358	**.162**	**.091**	**.176**	**.120**	**.798**	**.753**	**.095***
`RoBERTa-base`	Θ_T: MS-MARCO (msmarco-RoBERTa)	.095	.707	.662	.258	.359	.162	.087	.170	.116	.776	.732	.090
	Θ': Task Arithmetic ($\alpha = 1$)	.092	.700	.659	.250	.347	.156	.078	.156	.108	.784	.734	.089
	Θ': Task Arithmetic ($\alpha = 0.3$)	**.096**	**.720**	**.676**	**.259**	**.361**	**.165**	**.088**	**.173**	**.118**	**.806***	**.757***	**.093***
`DistilBERT`	Θ_T: MS-MARCO (msmarco-distilbert)	.093	.703	.662	**.258**	.357	.161	.087	.168	.115	.794	.744	.091
	Θ': Task Arithmetic ($\alpha = 1$)	.090	.689	.650	.249	.346	.156	.076	.147	.099	.710	.675	.083
	Θ': Task Arithmetic ($\alpha = 0.5$)	**.095**	**.720**	**.677**	.257	**.359**	**.163**	**.088**	**.171**	**.116**	**.806**	**.765**	**.094***

Table 1
Effectiveness of all models on Biomedical and Scientific domains. Best results are highlighted in boldface.

Table 1 presents the performance of our approach on the biomedical and scientific datasets, evaluated with five different models. In the initial evaluation, we fix $\alpha = 1$. Under this setting, Task Arithmetic outperforms the MS-MARCO fine-tuned baselines only on TREC-COVID with RoBERTa-base, T5-base, and T5-Large, and on SCIDOCS with T5-Large. These findings suggest the need for a small amount of labeled data to optimize the value of $alpha$. The remainder of this section focuses on the results obtained when α is optimized. Across all datasets and metrics, the Task Arithmetic–based model (Θ') consistently outperforms BM25, indicating its potential for effective domain adaptation in IR. For bi-encoders (DistilBERT and RoBERTa-base), Task Arithmetic yields consistent improvements over all baselines, with one exception: DistilBERT on the NFCorpus shows a minor drop in P@10

compared to the MS-MARCO fine-tuned counterpart (Θ_T). Nevertheless, our approach achieves a statistically significant improvement in MAP@100 on TREC-COVID, surpassing every baseline. For cross-encoders (T5-base and T5-Large), Task Arithmetic outperforms MonoT5 (Θ_T) on SCIDOCS and TREC-COVID, while producing comparable results on SciFact and NFCorpus. Notably, our method shows significant gains in MAP@100 on TREC-COVID, independent of the T5 variant, and yields statistically significant improvements in NDCG@10 on TREC-COVID and SCIDOCS. Regarding the decoder-only model (LLama-2), our approach (Θ') achieves superior performance on TREC-COVID compared to the MSMARCO fine-tuned version (Θ_T) and remains competitive on SCIDOCS. Table 2 extends the results to multilingual IR by using MT5-base as a cross-encoder for language-specific tasks. Both our proposed approach (Θ') and the IR-specific baseline (Θ_T) show statistically significant improvements over BM25 on each metric. Notably, our method surpasses the IR-specific model by up to 18% in NDCG@10, highlighting its effectiveness in adapting to new language settings. Thus, our approach successfully injects language-specific knowledge into the multilingual IR model (Θ_T), thereby enhancing its retrieval capabilities. These results further support Task Arithmetic as a lightweight but powerful strategy for zero-shot adaptation in multilingual IR. Finally, it is worth noting that the optimal scaling factor α exceeds 0.3 for all models and surpasses 0.7 for T5 variants and LLama-2, indicating that Task Arithmetic injects non-trivial domain knowledge into these retrieval models.

Model		GermanQuAD			MIRACL Spanish			MIRACL French			MIRACL English		
Re-ranker	Variant	P@10	NDCG@10	MAP@100	P@10	NDCG@10	MAP@100	P@10	NDCG@10	MAP@100	P@10	NDCG@10	MAP@100
BM25		.059	.437	.397	.135	.270	.215	.052	.174	.139	.107	.302	.247
MT5-base	Θ_T: MS-MARCO (en)	.069	.513	.463	.190	.379	.301	.071	.234	.186	.140	.398	.325
	Θ': Task Arithmetic ($\alpha = 1$)	**.071***	**.537***	**487***	**.200***	**.405***	**.325***	**.081***	**.278***	**.220***	**.151***	**435***	**.358***

Table 2
Effectiveness of all models on Language Transfer. Best results are highlighted in boldface.

5. Conclusion

In this paper, we investigate Task Arithmetic as a training-free method for zero-shot domain and language adaptation in IR, leveraging publicly available domain- and IR-specific LLMs. To this aim, we evaluate Task Arithmetic with six LLMs, including encoder-only, encoder-decoder and decoder-only architectures, across scientific, biomedical, and multilingual datasets. Our analysis shows that the proposed approach consistently improves the IR-specific model's performance across the board, reaching gains of up to 18% in NDCG@10.

6. Acknowledgments

We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support. This work was partially supported by the European Union – Next Generation EU within the project NRPP M4C2, Investment 1.,3 DD. 341 - 15 march 2022 – FAIR – Future Artificial Intelligence Research – Spoke 4 - PE00000013 - D53C22002380006, and by the MUR under the grant "Dipartimenti di Eccellenza 2023-2027" of the Department of Informatics, Systems and Communication of the University of Milano-Bicocca, Italy.

Declaration on Generative AI

During the preparation of this work, the authors used GPT3.5 and GPT-4 in order to: Grammar and spelling check, Paraphrase and reword. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Braga, P. Kasela, A. Raganato, G. Pasi, Investigating task arithmetic for zero-shot information retrieval, in: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 2738–2743. URL: https://doi.org/10.1145/3726302.3730216. doi:10.1145/3726302.3730216.

[2] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, D. Roth, Recent advances in natural language processing via large pre-trained language models: A survey, ACM Comput. Surv. 56 (2023). URL: https://doi.org/10.1145/3605943. doi:10.1145/3605943.

[3] A. Rogers, S. Luccioni, Position: Key claims in LLM research have a long tail of footnotes, in: Forty-first International Conference on Machine Learning, 2024.

[4] Y. Zhu, H. Yuan, S. Wang, J. Liu, W. Liu, C. Deng, H. Chen, Z. Liu, Z. Dou, J.-R. Wen, Large language models for information retrieval: A survey, arXiv preprint arXiv:2308.07107 (2023).

[5] Z. Qin, R. Jagerman, K. Hui, H. Zhuang, J. Wu, L. Yan, J. Shen, T. Liu, J. Liu, D. Metzler, X. Wang, M. Bendersky, Large language models are effective text rankers with pairwise ranking prompting, in: K. Duh, H. Gomez, S. Bethard (Eds.), Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 1504–1518. URL: https://aclanthology.org/2024.findings-naacl.97/. doi:10.18653/v1/2024.findings-naacl.97.

[6] P. Kasela, M. Braga, G. Pasi, R. Perego, Se-pqa: Personalized community question answering, in: Companion Proceedings of the ACM Web Conference 2024, WWW '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 1095–1098. URL: https://doi.org/10.1145/3589335.3651445. doi:10.1145/3589335.3651445.

[7] S. Zhuang, H. Zhuang, B. Koopman, G. Zuccon, A setwise approach for effective and highly efficient zero-shot ranking with large language models, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 38–47. URL: https://doi.org/10.1145/3626772.3657813. doi:10.1145/3626772.3657813.

[8] J. Lin, R. Nogueira, A. Yates, Pretrained transformers for text ranking: Bert and beyond, Springer Nature, 2022.

[9] E. Bassani, P. Kasela, G. Pasi, Denoising attention for query-aware user modeling, in: K. Duh, H. Gomez, S. Bethard (Eds.), Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 2368–2380. URL: https://aclanthology.org/2024.findings-naacl.153/. doi:10.18653/v1/2024.findings-naacl.153.

[10] X. Chen, X. Chen, B. He, T. Wen, L. Sun, Analyze, generate and refine: Query expansion with LLMs for zero-shot open-domain QA, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 11908–11922. URL: https://aclanthology.org/2024.findings-acl.708/. doi:10.18653/v1/2024.findings-acl.708.

[11] M. Li, H. Zhuang, K. Hui, Z. Qin, J. Lin, R. Jagerman, X. Wang, M. Bendersky, Can query expansion improve generalization of strong cross-encoder rankers?, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 2321–2326. URL: https://doi.org/10.1145/3626772.3657979. doi:10.1145/3626772.3657979.

[12] M. Braga, P. Kasela, A. Raganato, G. Pasi, Synthetic data generation with large language models for personalized community question answering, in: 2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2024, pp. 360–366. doi:10.1109/WI-IAT62293.2024.00057.

[13] T. Almeida, S. Matos, Exploring efficient zero-shot synthetic dataset generation for information retrieval, in: Y. Graham, M. Purver (Eds.), Findings of the Association for Computational Linguistics:

EACL 2024, Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 1214–1231. URL: https://aclanthology.org/2024.findings-eacl.81/.

[14] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[15] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, P. S. Yu, Generalizing to unseen domains: A survey on domain generalization, IEEE Transactions on Knowledge and Data Engineering 35 (2023) 8052–8072. doi:10.1109/TKDE.2022.3178128.

[16] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models, in: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.

[17] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, L. Deng, MS MARCO: A human generated machine reading comprehension dataset, in: T. R. Besold, A. Bordes, A. S. d'Avila Garcez, G. Wayne (Eds.), Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016, volume 1773 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2016. URL: https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf.

[18] E. Kamalloo, N. Thakur, C. Lassance, X. Ma, J.-H. Yang, J. Lin, Resources for brewing beir: Reproducible reference models and statistical analyses, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 1431–1440. URL: https://doi.org/10.1145/3626772.3657862. doi:10.1145/3626772.3657862.

[19] P. Kasela, G. Pasi, R. Perego, N. Tonellotto, Desire-me: Domain-enhanced supervised information retrieval using mixture-of-experts, in: European Conference on Information Retrieval, Springer, 2024, pp. 111–125.

[20] Y. Xia, J. Kim, Y. Chen, H. Ye, S. Kundu, C. C. Hao, N. Talati, Understanding the performance and estimating the cost of llm fine-tuning, in: 2024 IEEE International Symposium on Workload Characterization (IISWC), IEEE, 2024, pp. 210–223.

[21] M. Braga, Personalized large language models through parameter efficient fine-tuning techniques, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 3076. URL: https://doi.org/10.1145/3626772.3657657. doi:10.1145/3626772.3657657.

[22] G. Peikos, P. Kasela, G. Pasi, Leveraging large language models for medical information extraction and query generation, in: 2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2024, pp. 367–372. doi:10.1109/WI-IAT62293.2024.00058.

[23] G. Peikos, S. Symeonidis, P. Kasela, G. Pasi, Utilizing chatgpt to enhance clinical trial enrollment, arXiv preprint arXiv:2306.02077 (2023).

[24] E. Bassani, P. Kasela, A. Raganato, G. Pasi, A multi-domain benchmark for personalized search evaluation, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 3822–3827. URL: https://doi.org/10.1145/3511808.3557536. doi:10.1145/3511808.3557536.

[25] P. Kasela, G. Pasi, R. Perego, Park: Personalized academic retrieval with knowledge-graphs, Information Systems 134 (2025) 102574. URL: https://www.sciencedirect.com/science/article/pii/S0306437925000584. doi:https://doi.org/10.1016/j.is.2025.102574.

[26] G. Ilharco, M. T. Ribeiro, M. Wortsman, L. Schmidt, H. Hajishirzi, A. Farhadi, Editing models with

task arithmetic, in: The Eleventh International Conference on Learning Representations, 2022.
[27] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, et al., Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time, in: International conference on machine learning, PMLR, 2022, pp. 23965–23998.
[28] M. S. Matena, C. A. Raffel, Merging models with fisher-weighted averaging, Advances in Neural Information Processing Systems 35 (2022) 17703–17716.
[29] S. Ainsworth, J. Hayase, S. Srinivasa, Git re-basin: Merging models modulo permutation symmetries, in: The Eleventh International Conference on Learning Representations, 2023.
[30] L. Choshen, E. Venezian, N. Slonim, Y. Katz, Fusing finetuned models for better pretraining, arXiv preprint arXiv:2204.03044 (2022).
[31] M. Li, S. Gururangan, T. Dettmers, M. Lewis, T. Althoff, N. A. Smith, L. Zettlemoyer, Branch-train-merge: Embarrassingly parallel training of expert language models, in: First Workshop on Interpolation Regularizers and Beyond at NeurIPS 2022, 2022. URL: https://openreview.net/forum?id=SQgVgE2Sq4.
[32] J. Frankle, D. J. Schwab, A. S. Morcos, The early phase of neural network training, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=Hkl1iRNFwS.
[33] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for nlp, in: International conference on machine learning, PMLR, 2019, pp. 2790–2799.
[34] A. Chronopoulou, M. Peters, A. Fraser, J. Dodge, AdapterSoup: Weight averaging to improve generalization of pretrained language models, in: A. Vlachos, I. Augenstein (Eds.), Findings of the Association for Computational Linguistics: EACL 2023, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 2054–2063. URL: https://aclanthology.org/2023.findings-eacl.153/. doi:10.18653/v1/2023.findings-eacl.153.
[35] M. Braga, A. Raganato, G. Pasi, et al., Personalization in bert with adapter modules and topic modelling, in: Proceedings of the 13th Italian Information Retrieval Workshop (IIR 2023). Pisa, Italy, 2023, pp. 24–29.
[36] M. Braga, A. Raganato, G. Pasi, AdaKron: An adapter-based parameter efficient model tuning with kronecker product, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 350–357. URL: https://aclanthology.org/2024.lrec-main.32/.
[37] N. Daheim, N. Dziri, M. Sachan, I. Gurevych, E. Ponti, Elastic weight removal for faithful and abstractive dialogue generation, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 7096–7112. URL: https://aclanthology.org/2024.naacl-long.393/. doi:10.18653/v1/2024.naacl-long.393.
[38] A. Chronopoulou, J. Pfeiffer, J. Maynez, X. Wang, S. Ruder, P. Agrawal, Language and task arithmetic with parameter-efficient layers for zero-shot summarization, in: J. Sälevä, A. Owodunni (Eds.), Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024), Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 114–126. URL: https://aclanthology.org/2024.mrl-1.7/. doi:10.18653/v1/2024.mrl-1.7.
[39] F. Schlatt, M. Fröbe, M. Hagen, Lightning ir: Straightforward fine-tuning and inference of transformer-based language models for information retrieval, in: Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining, WSDM '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 1048–1051. URL: https://doi.org/10.1145/3701551.3704118. doi:10.1145/3701551.3704118.
[40] X. Ma, L. Wang, N. Yang, F. Wei, J. Lin, Fine-tuning llama for multi-stage text retrieval, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in

Information Retrieval, SIGIR '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 2421–2425. URL: https://doi.org/10.1145/3626772.3657951. doi:10.1145/3626772.3657951.

[41] E. Voorhees, T. Alam, S. Bedrick, D. Demner-Fushman, W. R. Hersh, K. Lo, K. Roberts, I. Soboroff, L. L. Wang, Trec-covid: constructing a pandemic information retrieval test collection, SIGIR Forum 54 (2021). URL: https://doi.org/10.1145/3451964.3451965. doi:10.1145/3451964.3451965.

[42] V. Boteva, D. Gholipour, A. Sokolov, S. Riezler, A full-text learning to rank dataset for medical information retrieval, in: Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38, Springer, 2016, pp. 716–722.

[43] A. Cohan, S. Feldman, I. Beltagy, D. Downey, D. S. Weld, Specter: Document-level representation learning using citation-informed transformers, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 2270–2282.

[44] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, H. Hajishirzi, Fact or fiction: Verifying scientific claims, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 7534–7550.

[45] T. Möller, J. Risch, M. Pietsch, Germanquad and germandpr: Improving non-english question answering and passage retrieval, in: Proceedings of the 3rd Workshop on Machine Reading for Question Answering, 2021, pp. 42–50.

[46] X. Zhang, N. Thakur, O. Ogundepo, E. Kamalloo, D. A. Hermelo, X. Li, Q. Liu, M. Rezagholizadeh, J. Lin, Miracl: A multilingual retrieval dataset covering 18 diverse languages, Transactions of the Association for Computational Linguistics 11 (2023) 1114–1131.

[47] V. Sanh, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter., in: Proceedings of Thirty-third Conference on Neural Information Processing Systems (NIPS2019), 2019.

[48] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. URL: https://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[49] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of machine learning research 21 (2020) 1–67.

[50] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, 2021. URL: https://arxiv.org/abs/2010.11934. arXiv:2010.11934.

[51] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. B. et al., Llama 2: Open foundation and fine-tuned chat models, 2023. URL: https://arxiv.org/abs/2307.09288. arXiv:2307.09288.

[52] O. Rohanian, M. Nouriborji, S. Kouchaki, F. Nooralahzadeh, L. Clifton, D. A. Clifton, Exploring the effectiveness of instruction tuning in biomedical language processing, Artificial Intelligence in Medicine 158 (2024) 103007. URL: https://www.sciencedirect.com/science/article/pii/S0933365724002495. doi:10.1016/j.artmed.2024.103007.

[53] L. N. Phan, J. T. Anibal, H. Tran, S. Chanana, E. Bahadroglu, A. Peltekian, G. Altan-Bonnet, Scifive: a text-to-text transformer model for biomedical literature, arXiv preprint arXiv:2106.03598 (2021).

[54] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. A. Smith, Don't stop pretraining: Adapt language models to domains and tasks, in: Proceedings of ACL, 2020.

[55] O. Rohanian, M. Nouriborji, S. Kouchaki, D. A. Clifton, On the effectiveness of compact biomedical transformers, Bioinformatics 39 (2023) btad103.

[56] R. Calizzano, M. Ostendorff, Q. Ruan, G. Rehm, Generating extended and multilingual summaries with pre-trained transformers, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 1640–1650. URL: https://aclanthology.org/2022.lrec-1.175/.

[57] L. Zhang, Y. Zhang, D. Long, P. Xie, M. Zhang, M. Zhang, Rankinggpt: Empowering large language models in text ranking with progressive enhancement, 2023. arXiv:2311.16720.

[58] R. Nogueira, Z. Jiang, R. Pradeep, J. Lin, Document ranking with a pretrained sequence-to-sequence model, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics:

EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 708–718. URL: https://aclanthology.org/2020.findings-emnlp.63/. doi:10.18653/v1/2020.findings-emnlp.63.

[59] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: http://arxiv.org/abs/1908.10084.

[60] L. H. Bonifacio, V. Jeronymo, H. Q. Abonizio, I. Campiotti, M. Fadaee, , R. Lotufo, R. Nogueira, mmarco: A multilingual version of the ms marco passage ranking dataset, 2021. arXiv:2108.13897.