POLITECNICO DI TORINO
Repository ISTITUZIONALE

Investigating Task Arithmetic for Zero-Shot Information Retrieval

Original
Investigating Task Arithmetic for Zero-Shot Information Retrieval / Braga, M., Kasela, P., Raganato, A., Pasi, G.. - (2025), pp. 2738-2743. (SIGIR '25: 48th International ACM SIGIR Conference on Research and Development in Information Retrieval Padua (ITA) July 13 - 18, 2025) [10.1145/3726302.3730216].

Availability:
This version is available at: 11583/3002210 since: 2025-07-29T13:42:39Z

Publisher:
Association for Computing Machinery

Published
DOI:10.1145/3726302.3730216

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

28 November 2026



Investigating Task Arithmetic for Zero-Shot Information Retrieval

Marco Braga
m.braga@campus.unimib.it
Department of Informatics, Systems and Communication - DISCo, University of Milano-Bicocca
Milano, Italy
DAUIN Dipartimento di Automatica e Informatica, Politecnico di Torino
Turin, Italy

Pranav Kasela
pranav.kasela@unimib.it
Department of Informatics, Systems and Communication - DISCo, University of Milano-Bicocca
Milano, Italy

Alessandro Raganato
alessandro.raganato@unimib.it
Department of Informatics, Systems and Communication - DISCo, University of Milano-Bicocca
Milano, Italy

Gabriella Pasi
gabriella.pasi@unimib.it
Department of Informatics, Systems and Communication - DISCo, University of Milano-Bicocca
Milano, Italy

Abstract

Large Language Models (LLMs) have shown impressive zero-shot performance across a variety of Natural Language Processing tasks, including document re-ranking. However, their effectiveness degrades on unseen tasks and domains, largely due to shifts in vocabulary and word distributions. In this paper, we investigate Task Arithmetic, a technique that combines the weights of LLMs pre-trained on different tasks or domains via simple mathematical operations, such as addition or subtraction, to adapt retrieval models without requiring additional fine-tuning. Our method is able to synthesize diverse tasks and domain knowledge into a single model, enabling effective zero-shot adaptation in different retrieval contexts. Extensive experiments on publicly available scientific, biomedical, and multilingual datasets show that our method improves state-of-the-art re-ranking performance by up to 18% in NDCG@10 and 15% in P@10. In addition to these empirical gains, our analysis provides insights into the strengths and limitations of Task Arithmetic as a practical strategy for zero-shot learning and model adaptation. We make our code publicly available at https://github.com/DetectiveMB/Task-Arithmetic-for-ZS-IR.

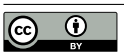
CCS Concepts

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

Keywords

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

ACM Reference Format:
Marco Braga, Pranav Kasela, Alessandro Raganato, and Gabriella Pasi. 2025. 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), July 13–18, 2025, Padua, Italy.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3726302.3730216



This work is licensed under a Creative Commons Attribution 4.0 International License.
SIGIR '25, July 13–18, 2025, Padua, Italy
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730216

1 Introduction

Large Language Models (LLMs) have achieved state-of-the-art performance in a wide range of tasks in several research fields [39, 49], including Information Retrieval (IR) [68]. By learning representations from massive unlabeled corpora, LLMs can be employed in document re-ranking [42, 45], query expansion [16, 34, 48], and synthetic data generation [3, 11]. Notably, these models often excel in zero-shot scenarios [1, 14, 59], effectively handling unseen tasks or domains without additional supervised fine-tuning. This zero-shot capability has driven their widespread use in document re-ranking [8, 35, 69], enabling robust retrieval even in the absence of domain-specific training data. Despite these advantages, domain mismatch remains a critical challenge that can significantly hamper the effectiveness of the model [58]. The BEIR benchmark [30, 54], which spans diverse tasks and domains, provides a heterogeneous framework for evaluating zero-shot IR performance. A common strategy is to pre-train a model on a large-scale IR dataset (e.g., MS-MARCO [41]) and then apply it zero-shot to unseen domains [30]. Although this approach often achieves strong performance, developing a single LLM that robustly generalizes to every domain remains very challenging [32, 54], as domain-specific adaptation typically requires large labeled datasets and considerable computational resources [61]. To mitigate these costs, Parameter-Efficient Fine-Tuning (PEFT) approaches have been proposed [12, 24, 25], which adapt a small subset of parameters while leaving most of the model frozen. Although PEFT significantly reduces the scale of updates and can be effective under limited data, labeled instances are still needed for each target domain, which makes it less suitable for frequent domain shifts or truly zero-shot scenarios. Meanwhile, the proliferation of publicly available LLMs, fine-tuned for diverse tasks, domains, and languages, on open-source platforms such as HuggingFace [29], offers a new opportunity to reuse existing models rather than training new ones from scratch. In this context, Task Arithmetic [27] emerges as a promising method. In fact, by merging the parameters of two or more fine-tuned LLMs through a simple addition or subtraction, Task Arithmetic transfers domain knowledge into a target model without further gradient-based optimization. More specifically, a Task Vector is defined as the difference

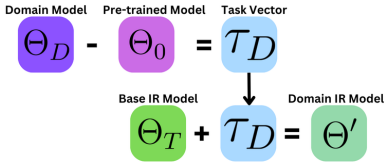


Figure 1: Proposed approach: Given a pre-trained LLM Θ_0 and its domain-finetuned version Θ_D, we compute the Task Vector τ_D as their parameter difference. To build a domain-specific IR model Θ', we add τ_D to an IR-finetuned model Θ_T.

between the parameters of a domain-fine-tuned model and its original pre-trained version, allowing domain-specific information to be injected or removed in parameter space with minimal overhead.

In this work, we investigate how Task Arithmetic can be applied to Information Retrieval, emphasizing domain and language transfer under zero-shot conditions. As shown in Figure 1, we begin with a pre-trained LLM and leverage domain-specialized vectors derived from fine-tuned models, combining them with an IR-focused baseline to produce a new domain-aware model. Our extensive evaluation covers eight publicly available datasets spanning scientific, biomedical, and multilingual tasks. We experiment with six LLMs with parameter counts ranging from 66 million to 7 billion, including encoder-only, encoder-decoder, and decoder-only architectures. Inasmuch as all specialized models are publicly available, our method remains reproducible, computationally efficient, and it does not require additional training thus minimizing carbon footprint. Our results show that Task Arithmetic consistently improves upon strong IR baselines, with gains of up to 18% in NDCG@10 and 15% in P@10, highlighting the practical value of reusing existing domain-trained models and underlying how Task Arithmetic is a lightweight but powerful strategy for zero-shot domain and language adaptation in IR.

2 Methodology

In this Section, we first present related work on model merging and introduce Task Arithmetic (Section 2.1). We then explain how we adapt this framework for zero-shot IR (Section 2.2).

2.1 Related Work and Motivation

Adapting LLMs to specialized domains typically requires resource-intensive fine-tuning. To address frequent domain shifts or label-scarce settings, recent studies explore weight interpolation and model merging [38, 60], leveraging the observation that models fine-tuned on related tasks may reside in compatible regions of parameter space [2, 17]. This allows arithmetic operations, such as averaging parameters, to preserve or even boost performance [22, 28, 33]. Among these methods, Task Arithmetic [27] is particularly appealing for the fact that it requires no training. Each domain- or task-specific fine-tuning run is represented as a Task Vector, defined by the difference between the fine-tuned model's parameters and those of the pre-trained model; adding or subtracting these vectors transfers knowledge without further updates [6, 18, 20, 26, 43]. Unlike approaches requiring adapters [10, 13, 24, 31] or gating modules [32, 37, 53], Task Arithmetic retains the original network architecture and remains cost-effective. In the context of zero-shot Information Retrieval (IR), Task Arithmetic remains relatively underexplored. Common approaches often rely on domain-specific adapters or fine-tuning [4, 5, 63–65], which are effective but still require training data and overhead. In contrast, Task Arithmetic simply reuses existing domain-fine-tuned LLMs, readily available on open-source platforms, and integrates them plug-and-play into an IR model. This approach is particularly appealing given the recent proliferation of LLMs fine-tuned for diverse domains and languages; by merging their specialized capabilities, we can construct a domain-aware IR model with minimal computational cost.

2.2 Task Arithmetic for Zero-Shot IR

In this Section we detail how Task Arithmetic can be applied in the context of Information Retrieval, particularly for domain and language transfer in zero-shot conditions. 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 [41]) 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:

$$\tau_D = \{\tau_1, \ldots, \tau_N\}, \quad \text{where} \quad \tau_i = (\theta_i)_D - (\theta_i)_0. \tag{1}$$

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:

$$\Theta' = \{\theta'_i = (\theta_i)_T + \alpha\tau_i\}_{i=1}^{N}. \tag{2}$$

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 a fully zero-shot scenario, which does not require additional labelled data, α is equal to one. If, instead, a small development set is available, the hyperparameter α can be optimized.

In summary, our framework involves three models: i) a publicly available pre-trained LLM, i.e. Θ_0, ii) the Θ_0 LLM fine-tuned on a specific domain, i.e. Θ_D, and iii) the Θ_0 LLM fine-tuned for IR, i.e. Θ_T. In many publicly released models, Θ_D is trained using a Language Modeling (LM) or Masked Language Modeling (MLM) objective, while Θ_T is fine-tuned to specific IR tasks like re-ranking or passage retrieval. As shown in Figure 1, we follow a three-step procedure: (1) **Task Vector Generation**: Using Equation 1, we compute τ_D by subtracting the pre-trained model's weights Θ_0 from the domain-fine-tuned model's weights Θ_D. This step captures the domain shift in parameters. (2) **Task Vector Integration**: Following Equation 2, we add the task vector τ_D (scaled by α) to the IR-specific model Θ_T, obtaining the Θ' model. This operation seamlessly transfers domain knowledge into the IR model without needing further backpropagation or labeled domain data. (3) **Zero-Shot Evaluation**: Finally, we directly evaluate the adapted model Θ' on IR tasks in the target domain. As no training steps are required, we can readily test multiple domains, languages, or tasks by simply substituting different Task vectors.

3 Experimental Setup

In this Section, we describe the datasets, evaluation metrics, and models used to assess the effectiveness of our proposed approach.

Model		SciFact				NFCorpus				SCIDOCS				TREC-COVID			
Re-ranker	Variant	P@10	NDCG@3	NDCG@10	MAP@100	P@10	NDCG@3	NDCG@10	MAP@100	P@10	NDCG@3	NDCG@10	MAP@100	P@10	NDCG@3	NDCG@10	MAP@100
BM25		.091	.637	.691	.649	.247	.404	.343	.154	.086	.156	.165	.112	.734	.764	.688	.085
Llama-2-7B	Θ_0: Pre-trained	**.099**	.701	.748	.702	.260	.431	.363	.165	**.099**	.179	**.191**	**.130**	.838	.845	.782	.095
	Θ_D: Domain-specific (MedTuned)	.097	.699	.742	.701	.250	.410	.349	.158	.095	.175	.184	.124	.812	.810	.761	.093
	Θ_T: MS-MARCO (RankingLlama)	**.099**	**.724**	**.770**	**.731**	**.265**	**.448**	**.373**	**.170**	.096	**.182**	.188	.129	.860	**.869**	.810	.098
	Θ': Task Arithmetic ($\alpha = 1$)	.096	.718	.757	.723	.265	.445	.370	.167	.095	.179	.185	.126	.858	.867	.801	.098
	Θ': Task Arithmetic (optimized $\alpha = 0.8$)	.097	**.728**	.765	.730	.262	.442	.365	.165	.097	**.182**	.189	.129	**.866**	.867	**.812**	**.099***
T5-Large	Θ_0: Pre-trained	.089	.639	.691	.650	.247	.404	.343	.154	.076	.144	.150	.102	.738	.757	.688	.086
	Θ_D: Domain-specific (SciFive)	.091	.637	.691	.649	.247	.404	.343	.154	.065	.111	.121	.082	.596	.582	.551	.076
	Θ_T: MS-MARCO (Mono-T5)	**.095**	**.706**	**.743**	**.709**	**.266***	**.431**	**.368***	**.167**	.095	.170	.182	.124	.784	805	.735	.092
	Θ': Task Arithmetic ($\alpha = 1$)	.092	.688	.721	.692	.257	.420	.356	.161	.096	**.176**	.185	.124	.816	**.818**	**.765**	.096
	Θ': Task Arithmetic ($\alpha = 0.9$)	.092	.699	.727	.699	.259	.423	.359	.162	**.098***	**.176**	**.187***	**.126**	**.818**	.805	.759	**.097***
T5-base	Θ_0: Pre-trained	.091	.638	.691	.649	.247	.404	.343	.154	.081	.152	.156	.105	.710	.761	.671	.084
	Θ_D: Domain-specific (SciFive)	.090	.638	.691	.648	.248	.400	.343	.154	.059	.110	.115	.080	.646	.662	.598	.079
	Θ_T: MS-MARCO (Mono-T5)	.096	.681	.726	.684	.258	**.424**	**.359**	**.162**	.090	.163	.173	.118	.762	.782	.712	.089
	Θ': Task Arithmetic ($\alpha = 1$)	.089	.645	.686	.649	.250	.405	.345	.156	.070	.131	.136	.094	.764	.825	.726	.088
	Θ': Task Arithmetic ($\alpha = 0.7$)	**.098**	**.702**	**.748**	**.708***	**.259**	**.424**	.358	**.162**	**.091**	**.171**	**.176**	**.120**	**.798**	**.836**	**.753**	**.095***
RoBERTa-base	Θ_0: Pre-trained	.090	.633	.686	.646	.248	.405	.344	.154	.072	.139	.142	.096	.612	.681	.579	.077
	Θ_D: Domain-specific (BioMed RoBERTa)	.091	.639	.691	.649	.241	.408	.340	.154	.071	.138	.140	.096	.600	.657	.561	.077
	Θ_T: MS-MARCO (msmarco-RoBERTa)	.095	.655	.707	.662	.258	.425	.359	.162	.087	.165	.170	.116	.776	.818	.732	.090
	Θ': Task Arithmetic ($\alpha = 1$)	.092	.649	.700	.659	.250	.408	.347	.156	.078	.153	.156	.108	.784	**.822**	.734	.089
	Θ': Task Arithmetic ($\alpha = 0.3$)	**.096**	**.669**	**.720**	**.676**	**.259**	**.432**	**.361**	**.165**	**.088**	**.169**	**.173**	**.118**	**.806***	.821	**.757***	**.093***
DistilBERT	Θ_0: Pre-trained	.091	.635	.689	.647	.246	.407	.345	.156	.082	.153	.159	.108	.712	.745	.666	.084
	Θ_D: Domain-specific (Bio-DistilBert)	.093	.661	.706	.663	.251	.405	.346	.155	.083	.151	.160	.108	.736	.789	.697	.086
	Θ_T: MS-MARCO (msmarco-distilbert)	.093	.657	.703	.662	**.258**	.418	.357	.161	.087	.159	.168	.115	.794	.808	.744	.091
	Θ': Task Arithmetic ($\alpha = 1$)	.090	.641	.689	.650	.249	.406	.346	.156	.076	.141	.147	.099	.710	.745	.675	.083
	Θ': Task Arithmetic ($\alpha = 0.5$)	**.095**	**.671**	**.720**	**.677**	.257	**.429**	**.359**	**.163**	**.088**	**.162**	**.171**	**.116**	**.806**	**.849**	**.765**	**.094***

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

The evaluation spans scientific, biomedical, and multilingual scenarios to show the potential of Task Arithmetic for zero-shot IR.

3.1 Datasets and Evaluation Metrics

We evaluate our approach on eight publicly available datasets across diverse domains. Four of these datasets are drawn from the BEIR benchmark [54]: TREC-COVID [56] and NFCorpus [9] address biomedical retrieval, while SCIDOCS [19] and SciFact [57] focus on scientific citation prediction and fact-checking, respectively. The remaining four datasets concern language-specific retrieval: GermanQuAD [40], which targets question answering in German, and three subsets (English, French, and Spanish) from the MIRACL multilingual IR challenge [67]. We focus on the biomedical and scientific domains, as they represent areas where both domain-specific and IR-focused models are available, thus enabling the application of Task Arithmetic. In contrast, we exclude BEIR datasets based on Wikipedia since the pre-trained language models used in our experiments have already been trained on Wikipedia [21, 36, 46, 55], and we cannot add domain-specific knowledge through Task Arithmetic. Retrieval effectiveness is measured using P@10, NDCG@3, NDCG@10 and MAP@100. Statistical significance is assessed via a Bonferroni-corrected two-sided paired student's t-test at 99% confidence. In all result tables, the symbol * indicates a statistically significant improvement over the best performing baseline.

3.2 Models and Baselines

For our experiments, we focus on two-stage retrieval and we use six different pre-trained language models (i.e. Θ_0) spanning multiple retrieval paradigms (bi-encoder, cross-encoder, LLM) and neural architectures (encoder-only, encoder-decoder, decoder-only). In details, we use `DistilBERT` [52] and `RoBERTa-base` [36] as encoder-only bi-encoders, `T5-base`, `T5-Large` [46], and `MT5-base` [62] as encoder-decoder cross-encoders, and `LLama-2-7b` [55] 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` [51], `SciFive` [44], `BioMed-RoBERTa` [23], and `Bio-DistilBERT` [50] for the biomedical and scientific domains, as well as `MT5-base-german`, `MT5-base-spanish`, `MT5-base-french`, and `MT5-base-english` [15] for language adaptations. These Task Vectors are then added to models fine-tuned on MS-MARCO (`RankingGPT-Llama2-7b` [66], `MonoT5` [42], `msmarco-RoBERTa` [47], `msmarco-distilbert` [47], and `MT5-base-msmarco` [7]), a widely adopted passage retrieval benchmark [41]. This setup facilitates direct comparisons with the same models specialized for either a specific language or domain (i.e. Θ_D), or IR (i.e. Θ_T). In a fully zero-shot scenario, we set $\alpha = 1$. Furthermore, in a setting where few labeled data are 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 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. we put $\alpha = 1$. We report both the results with the optimized and not-optimized α. All re-ranking experiments begin by retrieving the top 100 documents via BM25. Following a two-stage retrieval paradigm, 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, i.e. $\lambda_{BM25} = \lambda_{LLM} = 0.5$.

4 Results and Discussion

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 α. The

Model		GermanQuAD				MIRACL Spanish				MIRACL French				MIRACL English			
Re-ranker	Variant	P@10	NDCG@3	NDCG@10	MAP@100	P@10	NDCG@3	NDCG@10	MAP@100	P@10	NDCG@3	NDCG@10	MAP@100	P@10	NDCG@3	NDCG@10	MAP@100
BM25		.059	.381	.437	.397	.135	.248	.270	.215	.052	.125	.174	.139	.107	.251	.302	.247
MT5-base	Θ_0: Pre-trained	.050	.275	.336	.296	.094	.173	.191	.152	.035	.095	.127	.106	.075	.175	.212	.174
	Θ_D: Language specific	.051	.296	.353	.316	.097	.172	.191	.153	.045	.099	.143	.113	.079	.180	.225	.187
	Θ_T: MS-MARCO (en)	.069	.451	.513	.463	.190	.342	.379	.301	.071	.175	.234	.186	.140	.330	.398	.325
	Θ': Task Arithmetic ($\alpha = 1$)	**.071***	**.477***	**.537***	**487***	**.200***	**.373***	**.405***	**.325***	**.081***	**.215***	**.278***	**.220***	**.151***	**.366***	**435***	**.358***

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

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, on 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 all baselines and remains competitive on SCIDOCS. Interestingly, the pre-trained LLama-2 (Θ_0) attains the highest P@10 and NDCG@10 on SciFact and SCIDOCS, which likely reflects the extensive domain knowledge acquired during large-scale pre-training [55]. 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.

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) outperform all other baselines on each metric. Notably, our method shows statistically significant improvements over the IR-specific model by up to 18% in NDCG@10, highlighting its effectiveness in adapting to new language settings. The pre-trained MT5-base (Θ_0) and its language-specific variant (Θ_D) do not outperform BM25, suggesting that these models are not inherently optimized for IR tasks. In contrast, 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.

5 Ablation Study

We conduct an ablation study to examine scaling factor impact while applying Task Arithmetic. Table 3 presents NDCG@10 scores on NFCorpus and SciFact development sets for α values from 0.1 to 1.0 in increments of 0.1. The results reveal that retrieval performance varies considerably with changes in α. On SciFact, for instance, T5-base improves from .640 at $\alpha = 1.0$ to .722 at $\alpha = 0.7$, representing a notable difference of approximately 12%. In general, while values in the 0.7–0.9 range often yield strong results for models such as T5-base and T5-Large, no single α value is consistently optimal across all models or datasets. This variability underscores the importance of model-specific calibration. Moreover, $\alpha = 1.0$ rarely provides the best performance, suggesting that excessive emphasis on domain parameters can overshadow the IR-specific knowledge embedded in IR-tuned weights. RoBERTa-base and DistilBERT sometimes peak at moderate values between 0.3 and 0.5, whereas T5-based models and LLama-2 commonly favor higher settings. These observations indicate that a small grid search over α can be effective in identifying a tradeoff between the IR and the domain-specific knowledge.

Model	Dataset	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
LLama 2-7B	**SciFact**	.631	.643	.660	.675	.687	.705	**.712**	.711	.705	.704
	NFCorpus	.235	.238	.239	.241	**.242**	.241	.239	**.242**	.241	.241
T5-Large	**SciFact**	.728	.732	.727	.731	.730	.732	.728	.730	**.737**	.710
	NFCorpus	.307	.307	.307	.308	.308	.306	.306	**.311**	**.311**	.309
T5-base	**SciFact**	.712	.710	.712	.711	.709	.718	**.722**	.708	.673	.640
	NFCorpus	.302	.301	.304	.306	.306	.311	**.313**	.312	.303	.298
RoBERTa-base	**SciFact**	.713	.717	**.725**	.723	.722	.710	.715	.706	.695	.687
	NFCorpus	.334	.331	.332	.332	.333	**.334**	.330	.323	.315	.307
DistilBERT	**SciFact**	.723	.720	.723	.722	**.724**	.712	.705	.699	.688	.652
	NFCorpus	.334	**.336**	.333	.333	.333	.329	.325	.306	.297	.286

Table 3: Ablation study about the impact of the scaling factor.

6 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. These findings underscore Task Arithmetic as a lightweight yet powerful strategy for IR applications, particularly when computational resources or labeled data are limited.

Acknowledgment

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.

References

[1] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 1998–2022. https://doi.org/10.18653/v1/2022.emnlp-main.130

[2] Samuel Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. 2023. Git Re-Basin: Merging Models modulo Permutation Symmetries. In *The Eleventh International Conference on Learning Representations*.

[3] Tiago Almeida and Sérgio Matos. 2024. Exploring efficient zero-shot synthetic dataset generation for Information Retrieval. In *Findings of the Association for Computational Linguistics: EACL 2024*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 1214–1231. https://aclanthology.org/2024.findings-eacl.81/

[4] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 4623–4637. https://doi.org/10.18653/v1/2020.acl-main.421

[5] Elias Bassani, Pranav Kasela, Alessandro Raganato, and Gabriella Pasi. 2022. A multi-domain benchmark for personalized search evaluation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 3822–3827.

[6] Rishabh Bhardwaj, Duc Anh Do, and Soujanya Poria. 2024. Language Models are Homer Simpson! Safety Re-Alignment of Fine-tuned Language Models through Task Arithmetic. 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, 14138–14149. https://doi.org/10.18653/v1/2024.acl-long.762

[7] Luiz Henrique Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, , Roberto Lotufo, and Rodrigo Nogueira. 2021. mMARCO: A Multilingual Version of the MS MARCO Passage Ranking Dataset. arXiv:2108.13897 [cs.CL]

[8] Luís Borges, Rohan Jha, Jamie Callan, and Bruno Martins. 2024. Generalizable Tip-of-the-Tongue Retrieval with LLM Re-ranking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2437–2441. https://doi.org/10.1145/3626772.3657917

[9] Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. 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, 716–722.

[10] Marco Braga. 2024. 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* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 3076. https://doi.org/10.1145/3626772.3657657

[11] Marco Braga, Pranav Kasela, Alessandro Raganato, and Gabriella Pasi. 2024. 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)*.

[12] Marco Braga, Alessandro Raganato, and Gabriella Pasi. 2024. AdaKron: An Adapter-based Parameter Efficient Model Tuning with Kronecker Product. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 350–357. https://aclanthology.org/2024.lrec-main.32/

[13] Marco Braga, Alessandro Raganato, Gabriella Pasi, et al. 2023. Personalization in BERT with Adapter Modules and Topic Modelling. In *Proceedings of the 13th Italian Information Retrieval Workshop (IIR 2023). Pisa, Italy*. 24–29.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[15] Rémi Calizzano, Malte Ostendorff, Qian Ruan, and Georg Rehm. 2022. Generating Extended and Multilingual Summaries with Pre-trained Transformers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 1640–1650. https://aclanthology.org/2022.lrec-1.175/

[16] Xinran Chen, Xuanang Chen, Ben He, Tengfei Wen, and Le Sun. 2024. Analyze, Generate and Refine: Query Expansion with LLMs for Zero-Shot Open-Domain QA. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 11908–11922. https://doi.org/10.18653/v1/2024.findings-acl.708

[17] Leshem Choshen, Elad Venezian, Noam Slonim, and Yoav Katz. 2022. Fusing finetuned models for better pretraining. *arXiv preprint arXiv:2204.03044* (2022).

[18] Alexandra Chronopoulou, Jonas Pfeiffer, Joshua Maynez, Xinyi Wang, Sebastian Ruder, and Priyanka Agrawal. 2024. Language and Task Arithmetic with Parameter-Efficient Layers for Zero-Shot Summarization. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, Jonne Sälevä and Abraham Owodunni (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 114–126. https://doi.org/10.18653/v1/2024.mrl-1.7

[19] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2270–2282.

[20] Nico Daheim, Nouha Dziri, Mrinmaya Sachan, Iryna Gurevych, and Edoardo Ponti. 2024. Elastic Weight Removal for Faithful and Abstractive Dialogue Generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 7096–7112. https://doi.org/10.18653/v1/2024.naacl-long.393

[21] 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)*. 4171–4186.

[22] Jonathan Frankle, David J. Schwab, and Ari S. Morcos. 2020. The Early Phase of Neural Network Training. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Hkl1iRNFwS

[23] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of ACL*.

[24] 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 *International conference on machine learning*. PMLR, 2790–2799.

[25] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

[26] Shih-Cheng Huang, Pin-Zu Li, Yu-chi Hsu, Kuang-Ming Chen, Yu Tung Lin, Shih-Kai Hsiao, Richard Tsai, and Hung-yi Lee. 2024. Chat Vector: A Simple Approach to Equip LLMs with Instruction Following and Model Alignment in New Languages. 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, 10943–10959. https://doi.org/10.18653/v1/2024.acl-long.590

[27] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

[28] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*. Association For Uncertainty in Artificial Intelligence (AUAI), 876–885.

[29] Shashank Mohan Jain. 2022. Hugging face. In *Introduction to transformers for NLP: With the hugging face library and models to solve problems*. Springer, 51–67.

[30] Ehsan Kamalloo, Nandan Thakur, Carlos Lassance, Xueguang Ma, Jheng-Hong Yang, and Jimmy Lin. 2024. 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* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 1431–1440. https://doi.org/10.1145/3626772.3657862

[31] Pranav Kasela, Marco Braga, Gabriella Pasi, and Raffaele Perego. 2024. SE-PQA: Personalized Community Question Answering. In *Companion Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 1095–1098. https://doi.org/10.1145/3589335.3651445

[32] Pranav Kasela, Gabriella Pasi, Raffaele Perego, and Nicola Tonellotto. 2024. Desire-me: Domain-enhanced supervised information retrieval using mixture-of-experts. In *European Conference on Information Retrieval*. Springer, 111–125.

[33] Margaret Li, Suchin Gururangan, Tim Dettmers, Mike Lewis, Tim Althoff, Noah A. Smith, and Luke Zettlemoyer. 2022. Branch-Train-Merge: Embarrassingly Parallel Training of Expert Language Models. In *First Workshop on Interpolation Regularizers and Beyond at NeurIPS 2022*. https://openreview.net/forum?id=SQgVgE2Sq4

[34] Minghan Li, Honglei Zhuang, Kai Hui, Zhen Qin, Jimmy Lin, Rolf Jagerman, Xuanhui Wang, and Michael Bendersky. 2024. 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* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2321–2326. https://doi.org/10.1145/3626772.3657979

[35] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2022. *Pretrained transformers for text ranking: Bert and beyond*. Springer Nature.

[36] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL] https://arxiv.org/abs/1907.11692

[37] Yuning Mao, Lambert Mathias, Rui Hou, Amjad Almahairi, Hao Ma, Jiawei Han, Scott Yih, and Madian Khabsa. 2022. UniPELT: A Unified Framework for Parameter-Efficient Language Model Tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

[38] Michael S Matena and Colin A Raffel. 2022. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems* 35 (2022), 17703–17716.

[39] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. *ACM Comput. Surv.* 56, 2, Article 30 (Sept. 2023), 40 pages. https://doi.org/10.1145/3605943

[40] Timo Möller, Julian Risch, and Malte Pietsch. 2021. GermanQuAD and GermanDPR: Improving Non-English Question Answering and Passage Retrieval. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*. 42–50.

[41] 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 *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 (CEUR Workshop Proceedings, Vol. 1773)*, Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

[42] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 708–718. https://doi.org/10.18653/v1/2020.findings-emnlp.63

[43] Marinela Parović, Ivan Vulić, and Anna Korhonen. 2024. Investigating the Potential of Task Arithmetic for Cross-Lingual Transfer. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 124–137. https://aclanthology.org/2024.eacl-short.12/

[44] Long N Phan, James T Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature. *arXiv preprint arXiv:2106.03598* (2021).

[45] Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024. Large Language Models are Effective Text Rankers with Pairwise Ranking Prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 1504–1518. https://doi.org/10.18653/v1/2024.findings-naacl.97

[46] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.

[47] 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*. Association for Computational Linguistics. http://arxiv.org/abs/1908.10084

[48] Daniele Rizzo, Alessandro Raganato, and Marco Viviani. 2024. Comparatively Assessing Large Language Models for Query Expansion in Information Retrieval via Zero-Shot and Chain-of-Thought Prompting. In *Proceedings of the 14th Italian Information Retrieval Workshop, Udine, Italy, September 5-6, 2024 (CEUR Workshop Proceedings, Vol. 3802)*. CEUR-WS.org, 23–32. https://ceur-ws.org/Vol-3802/paper22.pdf

[49] Anna Rogers and Sasha Luccioni. 2024. Position: Key Claims in LLM Research Have a Long Tail of Footnotes. In *Forty-first International Conference on Machine Learning*.

[50] Omid Rohanian, Mohammadmahdi Nouriborji, Samaneh Kouchaki, and David A Clifton. 2023. On the effectiveness of compact biomedical transformers. *Bioinformatics* 39, 3 (2023), btad103.

[51] Omid Rohanian, Mohammadmahdi Nouriborji, Samaneh Kouchaki, Farhad Nooralahzadeh, Lei Clifton, and David A Clifton. 2024. Exploring the Effectiveness of Instruction Tuning in Biomedical Language Processing. *Artificial Intelligence in Medicine* 158 (2024), 103007. https://doi.org/10.1016/j.artmed.2024.103007

[52] V Sanh. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.. In *Proceedings of Thirty-third Conference on Neural Information Processing Systems (NIPS2019)*.

[53] Sainbayar Sukhbaatar, Olga Golovneva, Vasu Sharma, Hu Xu, Xi Victoria Lin, Baptiste Rozière, Jacob Kahn, Daniel Li, Wen-tau Yih, Jason Weston, et al. 2024. Branch-Train-MiX: Mixing Expert LLMs into a Mixture-of-Experts LLM. *arXiv preprint arXiv:2403.07816* (2024).

[54] 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 Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[55] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, and Yasmine Babaei et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL] https://arxiv.org/abs/2307.09288

[56] Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. TREC-COVID: constructing a pandemic information retrieval test collection. *SIGIR Forum* 54, 1, Article 1 (Feb. 2021), 12 pages. https://doi.org/10.1145/3451964.3451965

[57] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 7534–7550.

[58] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S. Yu. 2023. Generalizing to Unseen Domains: A Survey on Domain Generalization. *IEEE Transactions on Knowledge and Data Engineering* 35, 8 (2023), 8052–8072. https://doi.org/10.1109/TKDE.2022.3178128

[59] Shuai Wang, Harrisen Scells, Shengyao Zhuang, Martin Potthast, Bevan Koopman, and Guido Zuccon. 2024. Zero-shot Generative Large Language Models for Systematic Review Screening Automation. In *European Conference on Information Retrieval*. Springer, 403–420.

[60] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*. PMLR, 23965–23998.

[61] Yuchen Xia, Jiho Kim, Yuhan Chen, Haojie Ye, Souvik Kundu, Cong Callie Hao, and Nishil Talati. 2024. Understanding the performance and estimating the cost of llm fine-tuning. In *2024 IEEE International Symposium on Workload Characterization (IISWC)*. IEEE, 210–223.

[62] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. arXiv:2010.11934 [cs.CL] https://arxiv.org/abs/2010.11934

[63] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems* 36 (2023), 7093–7115.

[64] Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*.

[65] Jingtao Zhan, Qingyao Ai, Yiqun Liu, Jiaxin Mao, Xiaohui Xie, Min Zhang, and Shaoping Ma. 2022. Disentangled modeling of domain and relevance for adaptable dense retrieval. *arXiv preprint arXiv:2208.05753* (2022).

[66] Longhui Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, and Min Zhang. 2023. RankingGPT: Empowering Large Language Models in Text Ranking with Progressive Enhancement. arXiv:2311.16720 [cs.IR]

[67] Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics* 11 (2023), 1114–1131.

[68] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107* (2023).

[69] Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. 2024. 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* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 38–47. https://doi.org/10.1145/3626772.3657813