

TiKMiX: TAKE DATA INFLUENCE INTO DYNAMIC MIXTURE FOR LANGUAGE MODEL PRE-TRAINING (SUPPLEMENTARY MATERIAL)

Anonymous authors

Paper under double-blind review

EXPERIMENTAL SETUP

Datasets and Models Web data serves as one of the core sources for pre-training large language models (LLMs), playing a crucial role in enhancing model capabilities due to its broad coverage and diversity. However, precisely because web data encompasses a wide range of domains—including news, encyclopedias, forums, and academic content—its highly diverse origins make it extremely challenging to achieve a balanced mixture across different domains. We follow the same experimental setup as prior studies on web data mixture Wettig et al. (2025); Liu et al. (2025), utilize the RefinedWeb dataset Penedo et al. (2023), and employ the domain classifier He et al. (2023) to categorize the data into 26 distinct domains. Our models, ranging in size from 1B to 7B parameters, are trained on up to 1 trillion tokens. The training process is divided into two distinct stages, each consisting of 500 billion tokens, with a strategic adjustment of the data mixture ratio at the transition point between stages. We compare TiKMiX against several representative data mixing strategies: **Pile-CC** Gao et al. (2020): The original data mixture proposed by the authors of The Pile based on heuristics. **REGMiX** Liu et al. (2024): SOTA method that uses a regression model to predict and optimize validation loss for determining the mixture. **DoReMi** Xie et al. (2023): a classic dynamic data mixing method that relies on a proxy model. **QUAD** Zhang et al. (2025): a method for dynamic selection during training after clustering data. We use the best-reported mixture from their paper, re-normalized to the domains available in our setup.

Our proposed TiKMiX method achieves a balance between dynamic adaptability and computational efficiency in data mixture strategies. Similar to other dynamic approaches such as DoReMi and QUAD, TiKMiX adjusts the data mixture ratios according to the current state of the model. However, unlike these methods, TiKMiX does not require multiple iterations, which significantly improves training efficiency. Furthermore, TiKMiX simplifies the data mixing process and reduces engineering complexity without sacrificing model performance.

To systematically evaluate the effectiveness of different data mixing strategies, we conduct large-scale experiments on the RefinedWeb dataset. Our models range in size from 1B to 7B parameters and are trained on up to 1 trillion tokens. The training process is divided into two distinct stages, each consisting of 500 billion tokens. At the transition point between these two stages, we strategically adjust the data mixture ratios to further assess the impact of mixing strategies on model performance.

DOWNSTREAM TASK EVALUATION

To conduct a comprehensive and rigorous evaluation of our proposed method, we curated a diverse suite of nine widely-recognized downstream benchmarks. This evaluation matrix is strategically divided into two categories: **in-domain** and **out-of-domain**. This bifurcation allows for a dual-faceted assessment of our model’s capabilities: on one hand, to measure its proficiency on tasks closely aligned with its training objectives, and on the other, to critically examine its ability to generalize learned skills to novel tasks and knowledge domains. The consistent performance gains observed across both categories underscore our method’s ability to enhance the model’s foundational capabilities and foster robust generalization.

In-Domain Evaluation Our in-domain evaluation suite is designed to probe the model’s core competencies in complex reasoning, commonsense understanding, and knowledge-intensive appli-

cations. These benchmarks are thematically aligned with our method’s primary optimization goals and serve to quantify the depth of improvement in these critical areas.

- **MMLU (Massive Multitask Language Understanding)** Hendrycks et al. (2020): A highly challenging multitask benchmark that assesses knowledge across 57 disparate subjects, ranging from elementary mathematics and U.S. history to computer science and law. MMLU demands not only a vast repository of knowledge but also the ability to perform precise, domain-specific reasoning, making it a key indicator of a model’s comprehensive intellectual and academic capabilities.
- **HellaSwag** Zellers et al. (2019): A commonsense reasoning benchmark that tasks the model with selecting the most plausible continuation for a given context. HellaSwag is distinguished by its use of adversarially-generated distractors, which are designed to be highly confusable for models that rely on superficial statistical cues. It therefore serves as a robust test of a model’s deeper understanding of causality and everyday situations.
- **ARC (AI2 Reasoning Challenge)** Clark et al. (2018): This benchmark evaluates reasoning and comprehension on grade-school science questions. We assess performance on both its subsets: **ARC-Easy (ARC-E)**, which contains questions often solvable via information retrieval, and the more difficult **ARC-Challenge (ARC-C)**, which requires multi-step reasoning and synthesis of knowledge. Evaluating on both allows for a fine-grained analysis of the model’s capabilities, from basic knowledge retrieval to complex scientific inference.
- **TriviaQA** Joshi et al. (2017): A large-scale reading comprehension benchmark where questions are authored by trivia enthusiasts, leading to a high degree of diversity and complexity. The task requires models to locate answers within lengthy, evidence-rich documents, often amidst significant distractor information. It primarily evaluates the model’s proficiency in long-context processing, precise information retrieval, and fact verification.

Out-of-Domain Evaluation To rigorously assess the generalization power of our method, we selected a set of out-of-domain benchmarks that are distinct from the in-domain tasks in terms of subject matter, format, or required reasoning skills. Performance on these benchmarks directly reflects the model’s ability to transfer its learned meta-skills to new and unseen challenges.

- **PiQA (Physical Interaction QA)** Bisk et al. (2020): A commonsense benchmark focused on physical reasoning. Presented in a question-answering format, it requires the model to understand the properties and affordances of everyday objects (e.g., “How can you cool a cup of water faster?”). PiQA probes the model’s intuitive grasp of how the physical world operates, a domain of commonsense distinct from academic knowledge, making it an excellent test of generalization.
- **OpenBookQA** Mihaylov et al. (2018): This benchmark simulates an “open-book” exam, requiring the model to answer questions using a given set of elementary science facts. Success demands not only reading comprehension but, more importantly, the ability to reason over and combine these facts to answer questions whose solutions are not explicitly stated. It critically evaluates the model’s capacity for multi-step reasoning and knowledge application within a constrained context.
- **BoolQ (Boolean Questions)** Clark et al. (2019): A dataset of naturally occurring yes/no questions, sourced from real user search queries. The challenge lies in the fact that the relationship between the question and the provided evidence passage is often implicit, requiring sophisticated syntactic and semantic analysis to arrive at a correct Boolean judgment. BoolQ effectively measures the model’s fine-grained comprehension of natural, conversational language.
- **MathQA** Amini et al. (2019): A mathematical reasoning benchmark featuring multi-step word problems. The task requires models to parse natural language descriptions, formulate a correct sequence of operations, and execute them to find a solution. Covering a diverse range of mathematical reasoning categories, MathQA is a crucial benchmark for evaluating a model’s symbolic reasoning and logical chain-of-thought capabilities, representing a significant test of higher-order cognitive skills.

By systematically evaluating our method across this dual-category, nine-benchmark matrix, we demonstrate that our approach not only enhances performance in core competency areas (as shown

Table 1: Ablation study of REGMIX and TiKMiX on 1B and 7B models.

Benchmark	1B Model		7B Model	
	REGMIX	TiKMiX-D	REGMIX	TiKMiX-D
<i>In-Domain Benchmarks</i>				
MMLU Hendrycks et al. (2020)	31.5	32.2	40.7	41.5
HellaSwag Zellers et al. (2019)	56.0	57.4	76.6	76.4
ARC Easy Clark et al. (2018)	66.2	69.3	78.5	78.4
ARC Challenge Clark et al. (2018)	32.2	37.0	49.4	50.2
TriviaQA Joshi et al. (2017)	15.8	17.7	46.4	45.3
<i>Out-of-Domain Benchmarks</i>				
PiQA Bisk et al. (2020)	73.3	74.1	79.1	79.2
OpenBookQA Mihaylov et al. (2018)	37.0	37.4	43.2	45.4
MathQA Amini et al. (2019)	23.2	23.5	28.8	29.9
Average Perf.	43.9	45.5	55.3	56.0

by MMLU and ARC-C) but also significantly improves the transfer of these abilities to novel contexts (as evidenced by PiQA and MathQA). This comprehensive improvement across both in-domain and out-of-domain tasks provides strong evidence for the effectiveness and generalizability of our method.

To further investigate the impact of model scale on data utilization, we present a supplementary analysis in Figures 1 to 7. Our key finding is that models of different scales (1B and 7B) exhibit significantly different learning responses and form distinct preferences, even when trained on the exact same data. This phenomenon reveals a complex interplay between data utility and model scale. It provides a solid theoretical foundation for understanding and optimizing the data mixture for models of varying sizes.

EXPERIMENTS ON MODELS OF DIFFERENT SIZES

Considering computational overhead, for the 7B model, we adopted an experimental design similar to REGMIXLiu et al. (2024), training with 500B tokens in the first stage and 200B tokens in the second stage. Table 1 presents the experimental results of our method on models of different scales. It can be observed that our proposed method significantly outperforms the current state-of-the-art approach, REGMIX, on both in-domain and out-of-domain benchmarks. The performance on the 7B model effectively demonstrates the scalability of our approach. Furthermore, we note that unlike the 1B model, the 7B model’s performance on the benchmarks consistently improves throughout the training process. This suggests that the advantage of TiKMiX could be even more pronounced with additional training data.

OBSERVATION OF DATA MIXING WITH GROUP INFLUENCE

To conduct a rigorous analysis of inter-domain interactions during mixed training, we designed an experiment to test the principle of influence additivity. Our hypothesis was that the influence of a mixed dataset on a validation set could be accurately predicted by a weighted sum of the influences from its individual constituent domains. To verify this, we first established a baseline mixing recipe using our TiKMiX-D method. We then systematically explored the local space around this recipe by generating 256 perturbed configurations, created by applying a random scaling factor between 0.5 and 2.0 to each domain’s original proportion. After filtering out two sampling outliers, we proceeded with 254 unique data mixture configurations. For each of these 254 points, we sampled a corresponding 0.1B token dataset and measured its direct influence. We then compared this empirical influence value against a predicted influence, which was calculated by summing the pre-computed influences of each individual domain, weighted by their respective proportions in the mixture. As depicted in Fig 9, this comparison revealed a strong linear correlation. Specifically, the Pearson correlation coefficients on the ARCClark et al. (2018), HellaswagZellers et al. (2019), and TriviaQAJoshi et al. (2017) benchmarks reached 0.845, 0.848, and 0.931, respectively, all of which are statistically

highly significant ($p < 0.0001$). This result provides compelling evidence that the outcome of data mixing is highly predictable and can be modeled as a linear combination of inter-domain influences. Consequently, this finding offers a solid empirical justification for the theoretical soundness of our proposed two-stage optimization framework, encompassing both TiKMiX-D and TiKMiX-M.

STATEMENT ON THE USE OF LARGE LANGUAGE MODELS

LLMs were used solely for polishing the writing of this manuscript. All academic content, data analysis, and conclusions were independently produced by the authors. The role of LLMs was limited to improving the accuracy and fluency of the language.

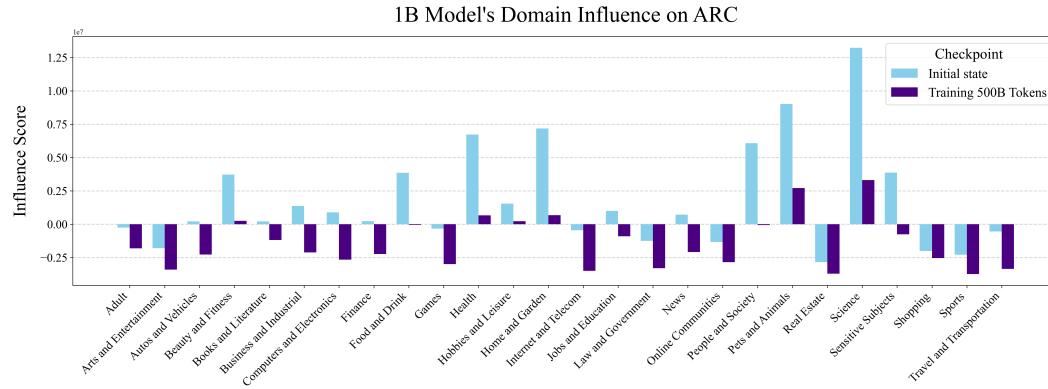


Figure 1: The impact of domains on a 1B model’s performance on the ARC benchmark as training progresses.

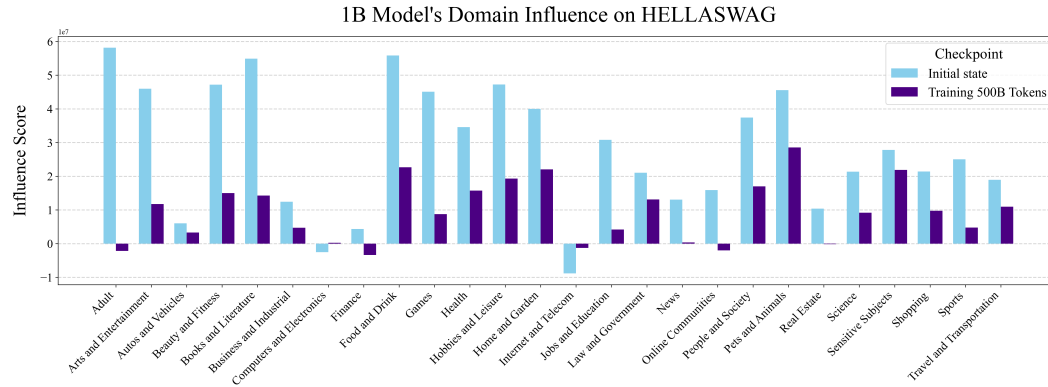


Figure 2: The impact of domains on a 1B model’s performance on the HELLASWAG benchmark as training progresses.

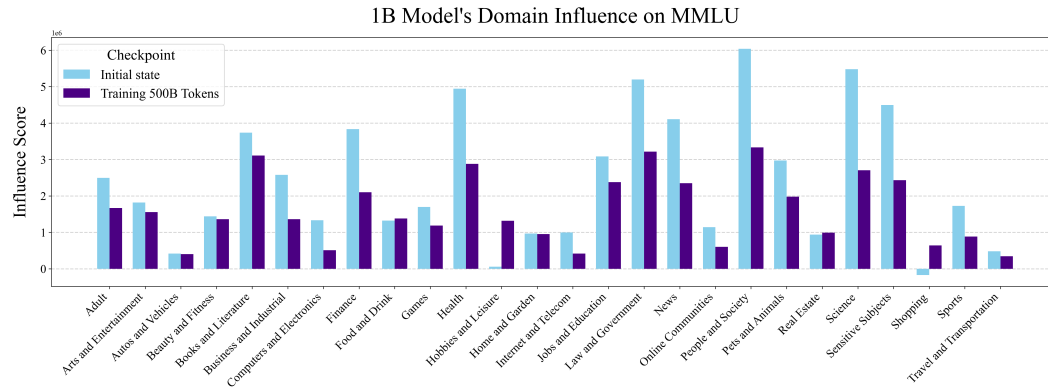


Figure 3: The impact of domains on a 1B model’s performance on the MMLU benchmark as training progresses.

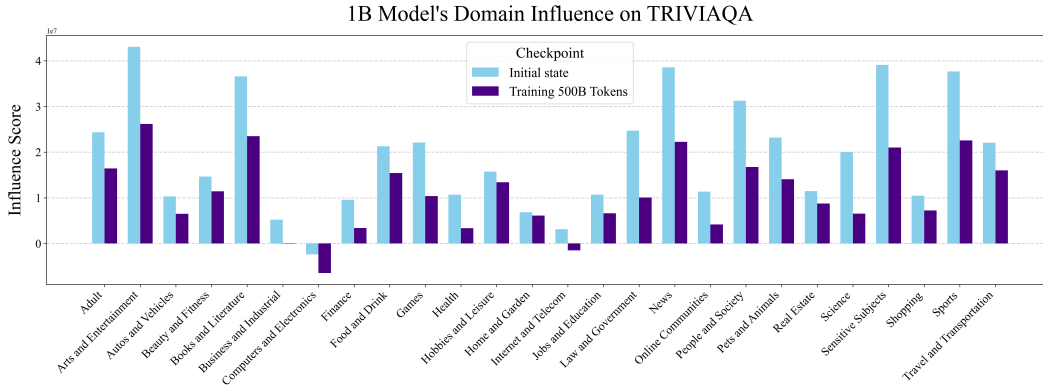


Figure 4: The impact of domains on a 1B model’s performance on the TRIVIAQA benchmark as training progresses.

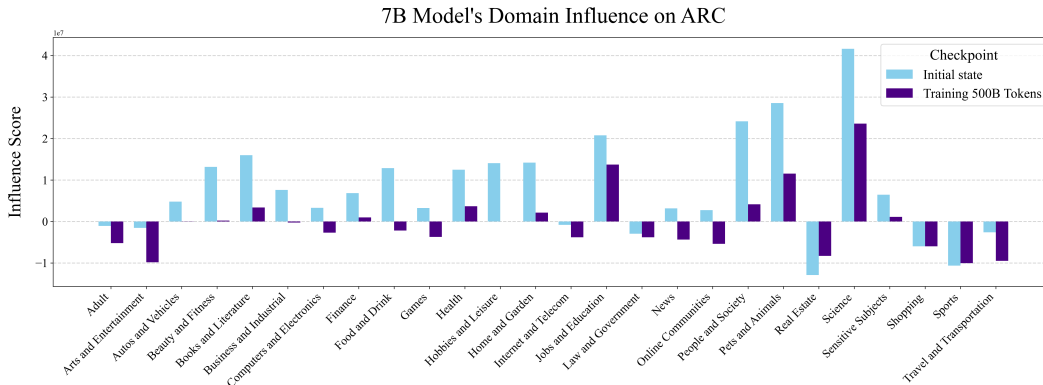


Figure 5: The impact of domains on a 7B model’s performance on the ARC benchmark as training progresses.

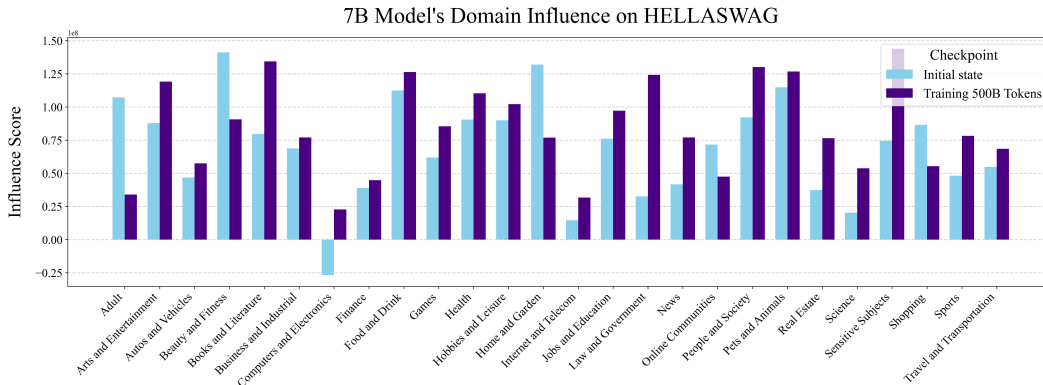


Figure 6: The impact of domains on a 7B model’s performance on the HELLASWAG benchmark as training progresses.

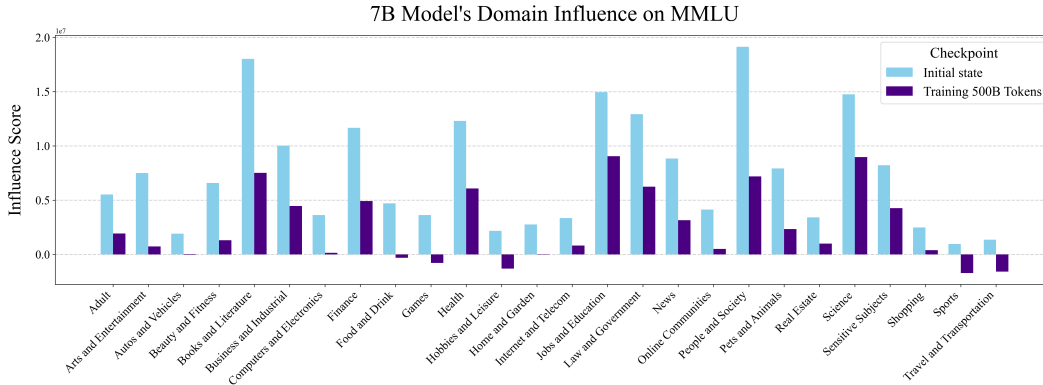


Figure 7: The impact of domains on a 7B model’s performance on the MMLU benchmark as training progresses.

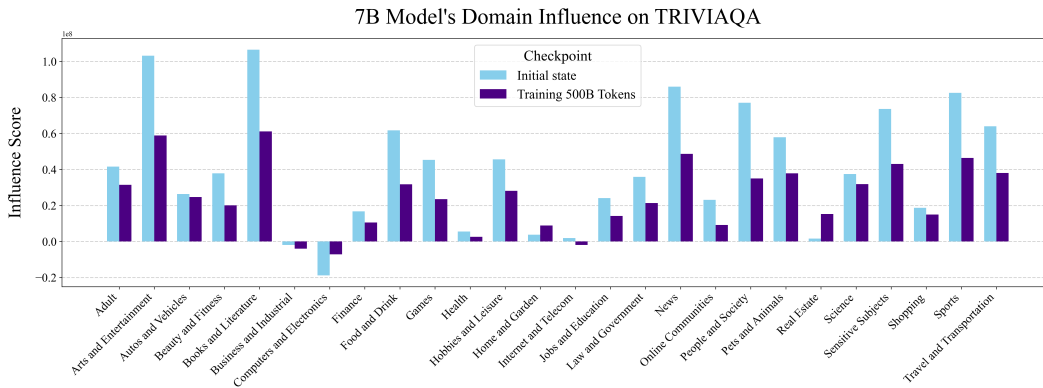


Figure 8: The impact of domains on a 7B model’s performance on the TRIVIAQA benchmark as training progresses.

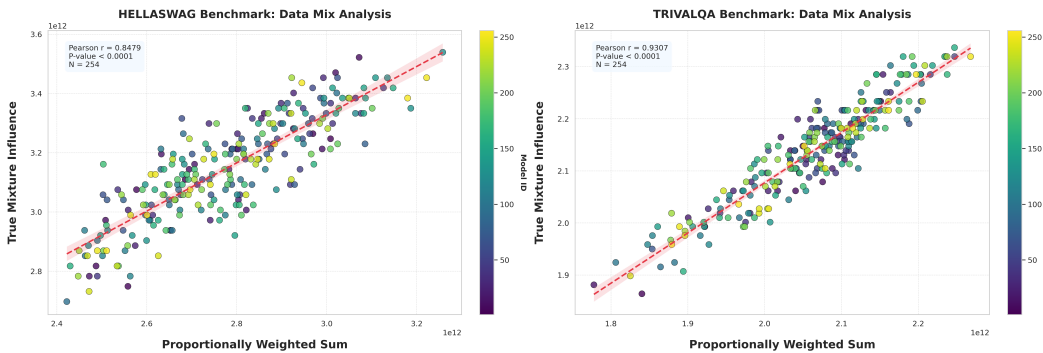


Figure 9: A Group Influence-based Analysis of Data Mixing Effects on Various Benchmarks.

REFERENCES

- Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. *arXiv preprint arXiv:1905.13319*, 2019.
- Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2023.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.
- Fengze Liu, Weidong Zhou, Binbin Liu, Zhimiao Yu, Yifan Zhang, Haobin Lin, Yifeng Yu, Bingni Zhang, Xiaohuan Zhou, Taifeng Wang, et al. Quadmix: Quality-diversity balanced data selection for efficient llm pretraining. *arXiv preprint arXiv:2504.16511*, 2025.
- Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. Regmix: Data mixture as regression for language model pre-training. *arXiv preprint arXiv:2407.01492*, 2024.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.
- Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refined-web dataset for falcon llm: Outperforming curated corpora with web data only. *Advances in Neural Information Processing Systems*, 36:79155–79172, 2023.
- Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. Organize the web: Constructing domains enhances pre-training data curation. *arXiv preprint arXiv:2502.10341*, 2025.
- Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems*, 36:69798–69818, 2023.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.
- Chi Zhang, Huaping Zhong, Kuan Zhang, Chengliang Chai, Rui Wang, Xinlin Zhuang, Tianyi Bai, Qiu Jiantao, Lei Cao, Ju Fan, et al. Harnessing diversity for important data selection in pretraining large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.