Lightweight Surrogate-Assisted Language Model Pretraining

Anonymous Authors[1]

Abstract

Standard causal language model pretraining uses a single-label cross-entropy objective that ignores the existence of multiple valid next-token continuations, resulting in sample inefficiency. Our initial analyses of language model outputs reveal that only a handful of tokens carry meaningful signal compared to the rest of the vocabulary. Predicated on this finding, we introduce a multi-label pretraining objective that modifies the loss to append a small set of context-conforming auxiliary tokens selected by a lightweight surrogate language model. Distinct from existing knowledge distillation methods, the surrogate is used only for token selection and suggestion rather than full distribution matching. We empirically show that adopting this multi-target stance achieves superior performance on benchmarks with notably less FLOPs utilization and tokens in our tested setting.

1. Introduction

It has been well established in recent years that scaling compute, model size, and training data can consistently improve model performance in next-token prediction and downstream tasks (Kaplan et al., 2020; Brown et al., 2020; Hoffmann et al., 2022). However, the act of scaling these subjects alone often incurs nontrivial costs, ultimately hindering faster training altogether. Incident to this, prevailing works have leveraged a teacher model to achieve enhanced model compression (Hinton et al., 2015; Gu et al., 2024), effectively reducing inference and training expenses.

Although knowledge distillation has been empirically proven to work in practice, inefficiencies regarding the teacher model's output distribution hinder further savings (Raman et al., 2023). More specifically, conventional causal language modeling assumes that a singular target is the only valid continuation for a given sequence. While it is possible to interpret knowledge distillation as a potential offset to this issue as it forces a base model to match a distribution that is not typically "one-hot", it ignores the fact that most of the reference distribution is noise (Table 1).

Multi-label learning (Bishop, 2006) also provides a relatively elegant solution, with most methods involving the notion that assigning singular labels to each sample is inherently near-sighted; that is, it is not uncommon for subjects to possess several plausible labels or in the language modeling case, several logical continuations. For instance, a single movie could belong to several genres or as shown in Figure 1, a sentence could be extended in multiple ways. Label smoothing (Szegedy et al., 2015; Müller et al., 2019) applied to a language model setting, similarly to knowledge distillation, could be viewed as a form of multi-label learning as it spreads the probability mass away from its previous singular instance. When applied as regularization while computing the cross-entropy loss, it encourages the model to raise the probabilities of predictions for non-target tokens. This achieves a "multi-label" stance in a nominal manner. However, explicit applications of multi-label learning to the LLM pretraining regime remain relatively untapped.

To analyze the aforementioned pitfalls of knowledge distillation, we swept over three language models of varying sizes and collated their average number of tokens produced in the output distribution above a specified probability threshold. Our findings show that only a small subset of tokens within a model's vocabulary carries a meaningful signal.

Motivated by this observation and the uncovered grounds of existing knowledge distillation and multi-label methods mentioned above, we propose lightweight surrogate-assisted language model pretraining (LSAP): the usage of cheap reference models to enhance pretraining efficiency via suggesting plausible targets. As shown in Figure 1, given an initial sequence and a target, we translate them to the surrogate's tokenizer using preconstructed mappings, pass the context and the translated version to the base and surrogate models respectively, and select tokens from the surrogate's output distribution that are assigned probabilities higher than a specified α. Those tokens are then appended into the base model's loss objective as negative log likelihoods.

We empirically show that LSAP induces efficiency gains

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

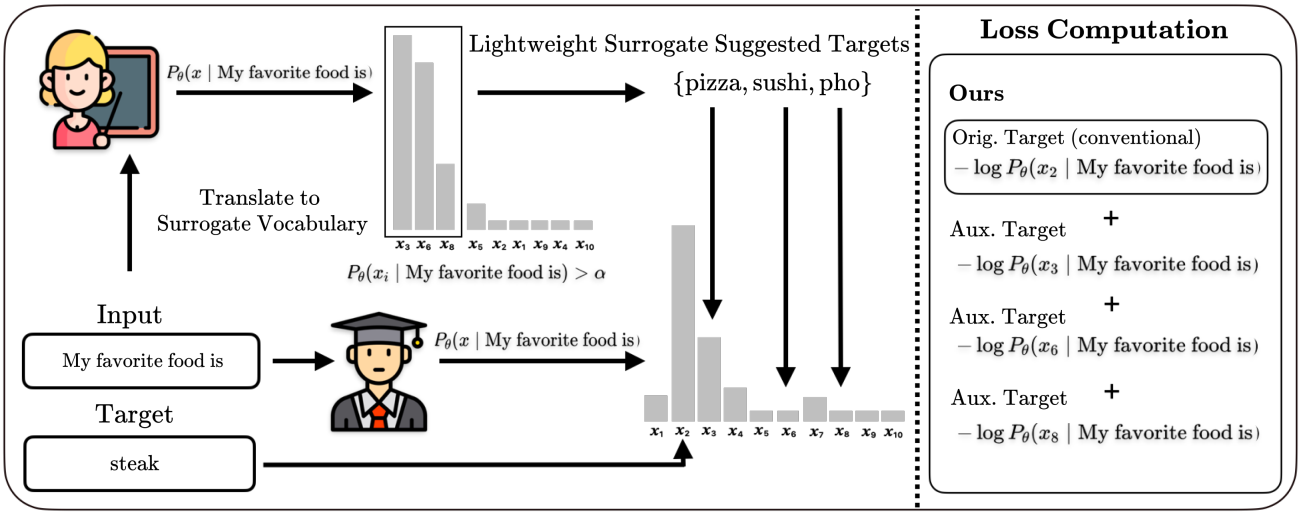

Figure 1. **Left:** A depiction of how additional target tokens are selected. **Right:** Our difference in loss function compared to typical cross entropy.

by pretraining OLMo2-1B from scratch on roughly 26B tokens from SlimPajama while leveraging Qwen3-0.6B as the reference LM. At identical training FLOPs, our method improves the average zero-shot benchmark by ∼3.23% over the standard cross-entropy baseline while utilizing 14.4% fewer training tokens. Similar gains were seen across HellaSwag, PIQA, LAMBADA, ARC-Easy, ARC-Challenge, OpenBookQA, and WinoGrande. Remarkably, challenging the standard knowledge-distillation assumption that teacher language models should be larger than the student, LSAP shows that the converse possesses tangible benefits.

2. Related Works

Pretraining Data Optimization The promise of a reduced carbon footprint and minimized expenses largely contributes to the ever-increasing usage of data selection algorithms. Methods aimed towards optimizing datasets have been empirically proven to strengthen language model performance, with many targeting distribution matching (Brown et al., 2020; Kang et al., 2024), quality (Li et al., 2024; Du et al., 2022; Wenzek et al., 2019), and diversification (Wei et al., 2022). In the pretraining regime, previous works have established the notion of scoring and selecting samples with the lowest uncertainties (Coleman et al., 2020). In the same vein, perplexity-based methods are also used to prune data provided a high quality n-gram model (Wenzek et al., 2019; Ankner et al., 2025; Laurençon et al., 2023; Muennighoff et al., 2023). However, the scoring process is typically storage-intensive and using a predefined uncertainty or perplexity threshold can incur the selection of harmful tokens via a large threshold or prevent beneficial tokens from being selected through a threshold too low.

Knowledge Distillation (KD) Hinton et al. (2015) first demonstrated that training a small network to mimic a powerful teacher model's predictions could recover a large portion of the teacher's performance with a significantly smaller model. Introduced into the causal language modeling setting by Kim & Rush (2016), many prevailing works have employed this strategy to achieve enhanced model compression (Gu et al., 2024; Datta et al., 2025) and continual learning (Li & Hoiem, 2018; Zhang et al., 2025a). This is typically performed by minimizing the forward Kullback-Leibler divergence between the output distributions of the teacher and student models. However, it has been recently shown that the teacher's output distributions are usually peaked, ultimately making tail regions of low probabilities contribute negligibly to the KL sum (Raman et al., 2023; Dasgupta et al., 2026). Contemporary fixes to this either reweigh the KL divergence to amplify tail tokens (Dasgupta et al., 2026) or retain a small subset of logit values from the teacher (Raman et al., 2023; Wang & Zhou, 2025). Although small language models have been used to aid pretraining through KD (Rawat et al., 2026), they still utilize distribution matching and implement stringent logit filters. By contrast, our method sidesteps this entirely by using a lightweight surrogate model solely for the selection of additional targets during pretraining. By using a probability threshold as opposed to a predefined hyperparameter that retains top-k tokens, we can dynamically scale the number of auxiliary targets selected per position by the teacher model.

Multi-Label Learning Most supervised learning methods assume that data is partitioned cleanly into unique and independent classes (Bishop, 2006). However, this may not always be the case. There often exist multiple categories in which data can be assigned to (e.g. a movie belonging

Table 1. **Average number of tokens above a probability threshold α.** All models are evaluated on 5000 examples from each dataset. Each entry is averaged over seeds 21, 42, and 67. Here, WikiText (WT), OpenWebText (OWT), and FineWeb-Edu (FW-E) have been abbreviated for brevity and textual coherence.

Model	α	WT	OWT	FW-E	ArXiv	C4	IMDb	PG19	AVG
Qwen3-0.6B	0.01	9.26	9.77	9.37	9.54	9.91	10.37	8.94	9.59
Qwen3-0.6B	0.05	2.77	2.85	2.84	2.60	2.86	2.81	2.39	2.73
Qwen3-0.6B	0.2	0.87	0.86	0.91	0.83	0.84	0.78	0.81	0.84
Llama-3.1-8B	0.01	8.61	8.82	8.69	9.16	8.81	9.89	6.83	8.69
Llama-3.1-8B	0.05	2.77	2.85	2.84	2.74	2.79	2.94	2.29	3.14
Llama-3.1-8B	0.2	0.99	1.00	1.02	0.94	0.98	0.92	0.99	0.98
Mistral-7B-v0.3	0.01	8.33	7.94	8.02	9.01	8.51	9.32	6.32	8.21
Mistral-7B-v0.3	0.05	2.75	2.68	2.72	2.71	2.78	2.84	2.14	2.66
Mistral-7B-v0.3	0.2	0.98	1.03	1.04	0.95	1.00	0.95	1.03	1.00

to several genres). Against this background, multi-label objectives have emerged in which a model attempts to learn from a set of examples where each possesses a set of plausible labels or classes (Tsoumakas & Katakis, 2009). Initial applications have been seen in pattern recognition (Elisseeff & Weston, 2001), text categorization (Schapire & Singer, 2000) and medical diagnosis (Pestian et al., 2007). Multi label classification tasks have also been studied in the natural language processing setting, chiefly with the observation that output token distributions exhibit lower entropy as we increase model scale (Ma et al., 2025). This also suggests that when using a larger teacher model, less tokens provide meaningful signal to the student. Our method points out this single-label inefficiency and attempts to alleviate it within the pretraining regime. As far as we know, no prior works have introduced an auxiliary loss function wherein the pretraining objective is inherently multi-label through sampling from the output distributions of a lightweight surrogate.

3. Methodology

Empirical Motivations Causal language modeling approaches leverage a categorical cross entropy loss function. To allow for the dynamic selection of additional and mainly *helpful* tokens, we assume that there exists a performant surrogate model that produces coherent text.

It is known that in knowledge distillation, not all tokens within the teacher's output distribution contributes meaningful signal to the student. To obtain a deeper analysis, we ran inference on three models across several datasets and computed the average number of tokens that were assigned probabilities higher than α in the output distributions. As shown in Table 1, when α is fixed to a value as low as 0.01, only roughly 8.83 tokens were assigned probabilities higher than that. Having most of the probability mass concentrated within a small pool of tokens suggests that the Kullback-Leibler divergence computed between the teacher and student's output distributions within KD is computationally inefficient. More specifically, it would require summing over thousands of tokens when most tail probabilities contribute negligibly to the KL value. Notably, with a probability threshold as permissive as 0.01, only roughly 9.59 out of tens of thousands were selected on average.

However, when selecting additional tokens based on fixed probability thresholds, a natural issue would be including incorrect targets. Choosing an α too small could induce the selection of incorrect targets while an α too large could yield nominal gains with respect to the surrogate's inference cost. Acknowledging this, we compare three α values in Section 4 and record their respective benchmark results.

Causal Language Modeling To begin, let $\mathcal{T}_b : \Sigma_b \to V_b$ denote the base model's tokenizer where Σ_b and V_b represents the set of token strings and the token ID's respectively. Provided an input sequence $x = (x_1, x_2, \ldots, x_N)$ where $x_i \in V_b^*$ standard causal language modeling paradigms employ the following cross-entropy loss:

$$\mathcal{L}_{CLM}(\theta; x) = -\frac{1}{N}\sum_{i=1}^{N} \log P_\theta(x_i | x_{<i})$$

In this setting, $L_{CLM}(\theta; x)$ denotes the loss function parameterized by θ with an input sequence $x \in \mathcal{X}$ while $x_{<i}$ refers to x_i's preceding tokens.

Multi-Label Objective Unlike conventional logit-based knowledge distillation (Hinton et al., 2015), our proposal alters the loss function by appending additional loss objectives (i.e., pure negative log likelihoods rather than the KL-divergence) using a surrogate model $\mathcal{S}$. However, there may exist discrepancies in tokenization. For instance, the word "because" may be represented as a single token within the base model's vocabulary while split into "be" + "cause" for the surrogate's. To mitigate this issue, we only use tokens that exist in the intersection between both vocabularies. With $\mathcal{T}_s : \Sigma_s \to V_s$ being the surrogate model's tokenizer,

let Σ_b and Σ_s denote the token string sets of $\mathcal{T}_b$ and $\mathcal{T}_s$ respectively. Then, $\Sigma_\cap := \Sigma_b \cap \Sigma_s$. We then establish translation mappings $\mathcal{F}_{b\to s} : \mathcal{T}_b(\Sigma_\cap) \to \mathcal{T}_s(V_s)$ for encoding inputs and $\mathcal{F}_{s\to b} : \mathcal{T}_s(\Sigma_\cap) \to \mathcal{T}_b(\Sigma_\cap)$ for decoding logits. Then, we can apply the following filters to the inputs and outputs to mitigate tokenizer discrepancies:

$$\mathcal{F}_{b\to s}(x_i) = \begin{cases} \mathcal{T}_{b\to s}(x_i) & \text{if } x_i \in \mathcal{T}_b(\Sigma_\cap) \\ -100 & \text{if } x_i \notin \mathcal{T}_b(\Sigma_\cap) \end{cases}$$

$$\mathcal{F}_{s\to b}(v_i) = \begin{cases} \mathcal{T}_{s\to b}(v_i) & \text{if } v_i \in \mathcal{T}_s(\Sigma_\cap) \\ -\infty & \text{otherwise or token id = -100} \end{cases}$$

Here, $v_i \in V_s$ denotes surrogate token indices. In our experiments, -100 denotes the padding token id for the surrogate model's input mask, which forces tokens not existing within $\Sigma_\cap$ to be ignored by the attention mechanism. This imposes a constraint upon our proposal: the cardinality of $\Sigma_\cap$ should be similar to that of Σ_b in order to retain input quality when passed to the surrogate.

For a provided target token x_i and preceding tokens $x_{<i}$, we use the $\mathcal{S}$'s output logits to select the top-k tokens that fall above a probability threshold $a \in [0, 1]$. To this end, passing the translated sequence $\mathcal{F}_{b\to s}(x_{<i})$ into $\mathcal{S}$ yields an output distribution over V_s. We set the probability of the target $\mathcal{F}_{b\to s}(x_i)$ and tokens not in $\Sigma_\cap$ to $-\infty$ to avoid selection. For a given target and an array of surrogate target indices v, the auxiliary loss objective can be formalized in the following regard:

$$\mathcal{L}_{surrogate} = -\sum_{i=0}^{|v|} \log P_\theta(\mathcal{F}_{s\to b}(v_i)|x_{<i})$$

Here, $|v|$ denotes the number of auxiliary tokens selected for a given target while $\mathcal{F}_{s\to b}(v_i)$ denotes a target selected by the surrogate translated to the base model's vocabulary. With this, we also acknowledge the fact that the validity of the surrogate's outputs are crucial to our method's success. If the reference model delegates high probabilities to tokens that are factually incorrect, the base model will leverage those indices nonetheless. To partially offset this issue, we multiply the auxiliary loss by a scaling factor following a cosine-annealing scheduler λ to achieve the combined loss:

$$\mathcal{L}_{ours}(\theta; x) = \frac{1}{N \cdot |v|}\left[\mathcal{L}_{CLM} + \lambda(t) \cdot \mathcal{L}_{surrogate}\right]$$

For this specific formulation, $\mathcal{L}_{CLM}$ and $\mathcal{L}_{surrogate}$ are unaveraged as the outside $1/(N \cdot |v|)$ accounts for both loss terms. Intuitively, the scheduler would prevent contaminating downstream training inputs with potentially false objectives. This allows the data distribution to dominate during later stages of training where accuracy is more important with respect to downstream performance (Zhang et al., 2025b).

4. Experiments

4.1. Setup

Baseline Setting For our baseline, we train OLMo2-1B (OLMo et al., 2025) on ∼26 billion tokens from SlimPajama (Soboleva, 2023).With AdamW as the optimizer, we set the maximum sequence length to 1024 with a maximum learning rate at 4e-4 with a cosine decaying scheduler.

LSAP Setup Past work has found that using the instruct version of LMs as a teacher for knowledge distillation in a pretraining setting yields higher scores in benchmarks compared to their pretrained or "base" counterparts (Liu et al., 2026). For this reason, we mainly use the post-trained variants of our reference models. In our experiments, we leverage Qwen3-0.6B (Team, 2025) as our surrogate while pretraining OLMo2-1B from scratch. Additionally, the tokenizers used in both models contain highly similar tokens, further motivating its use. Based on our observations in Table 1, we extract 15 tokens with the highest probability from the surrogate's output distribution per position and filter the tokens according to our chosen α threshold of 0.05. Aside from this, we use the same training configuration as listed in **Baseline Setting**. Due to our relatively minimal dataset size, we did not find it necessary to use the cosine annealing scheduler for the surrogate loss term. However, we still included its results with our best α.

Evaluation Setup To evaluate the effectiveness of our approach, we compare benchmark performance differences between a model trained with typical cross entropy and LSAP. To this end, we employ the lm-evaluation-harness (Gao et al., 2024).

4.2. Results

Here, we compare standard next-token pretraining against LSAP through matched FLOPs comparisons. All comparisons with LSAP includes the FLOPs overhead of the surrogate's inference.

Notably, we find that our method achieves higher benchmark performance at matched TFLOPs. As shown in Figure 2, LSAP with α = 0.05 and 0.05 without annealing achieves approximately 3.23% higher scores compared to the baseline at a matched 232 million training TFLOPs. Interestingly, our run with α = 0.05 generally outperformed runs with

Table 2. **Performance across benchmarks.** We compare the benchmark scores of LSAP to the baseline at matched TFLOPs (232 million). HellaSwag (HeSw), LAMBADA (LMB), ARC-Easy (ARC-E), ARC-Challenge (ARC-C), WinoGrande (WinoG), OpenBookQA (OBQA), and TruthfulQA (TrQA) are abbreviated for stylistic simplicity. The best result over each benchmark is bolded.

Loss	α	HeSw	PIQA	LMB	ARC-E	ARC-C	WinoG	OBQA	TrQA
LSAP	0.05-anneal	**47.40**	**69.40**	**32.40**	46.60	**26.00**	56.40	32.20	38.69
LSAP	0.05	**47.40**	66.80	30.60	**47.80**	23.40	**57.40**	**33.00**	39.83
LSAP	0.01	45.40	66.80	27.20	47.00	23.20	55.20	29.40	40.30
LSAP	0.20	45.40	67.40	28.00	44.40	24.60	53.60	29.60	**40.70**
Cross Entropy	0	44.09	65.82	26.38	44.03	22.89	51.01	29.49	40.12

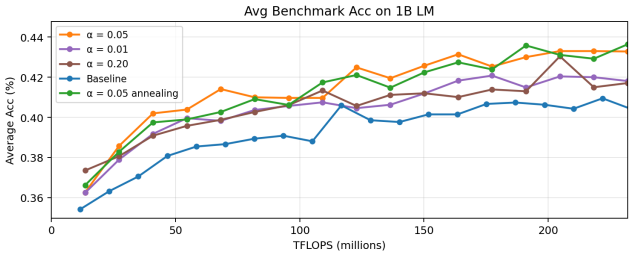

Figure 2. **Average benchmark scores.** We compare the mean zero-shot benchmark scores of LSAP against the baseline while varying α.

$\alpha = 0.01$ and 0.20, implying that a threshold of 0.01 may have been too permissive to the point of including false targets while 0.2 was overly restrictive and did not provide enough guidance.

Additionally, we note that applying the surrogate cosine annealing scheduler did not induce significant performance gains in our tested setting. We attribute this to our scope of training. Pretraining involves the task of broad knowledge acquisition while later stages (i.e., mid-training and post-training) require more factual suggestions that a small surrogate language model cannot provide consistently. We did not extend our experiments into those regions and thus leave it as a supplemental solution for future extensions.

LSAP generally outperforms our baseline in benchmark accuracy at matched FLOPs. Compared to the cross entropy loss, LSAP with $\alpha = 0.05$ achieves +3.31% on HellaSwag, +3.58% on PIQA, +6.02% on LAMBADA, +2.57% on ARC-Easy, +3.11% on ARC-Challenge, +5.39% on WinoGrande, +2.71% on OpenBookQA, and -1.43% on TruthfulQA.

Given that LSAP achieves higher benchmark performance while using 14.4% fewer tokens at equivalent compute expenses, this suggests improved sample efficiency rather than gains driven purely by additional compute.

5. Limitations and Future Work

Due to a lack of proper compute resources, we were ultimately unable to apply LSAP to models of varying architecture and size. However, a natural extension of this work would be to analyze how we can make the surrogate prior to significant benchmark degradations. Additionally, we are still unsure how essential having a high tokenizer intersection percentage is to LSAP's efficacy. This is easily verifiable through a separate run wherein specific tokens are zeroed out to artificially decrease the intersection proportion.

Although it has not been empirically proven, it is intuitive to think that the benefits of LSAP should scale proportionally with the size of the base model while keeping the surrogate's size constant. In this scenario, the surrogate would provide the same target suggestions regardless of the base model's size. As a result, significantly scaling up the size of the base model would decrease the relative FLOPs overhead of the surrogate's inference.

6. Conclusion

In this work, we propose LSAP, a lightweight surrogate-assisted pretraining method that addresses the single-label inefficiency present in conventional causal language modeling. Our initial study of token distributions across multiple models and datasets reveal that only a small fraction of tokens carry meaningful signal, motivating our usage of a lightweight reference model to suggest additional training targets. By appending auxiliary tokens as negative log-likelihoods rather than performing full distribution matching, we avoid the computational overhead of conventional knowledge distillation while still benefiting from a multi-label objective. Contrasting with the standing view that a larger reference model is needed for enhanced results, our experiments on OLMo2-1B show that LSAP achieves superior benchmark performance using fewer tokens and matched FLOPs, suggesting improved sample efficiency while using a surrogate LM that is smaller than the one we are pretraining.

Impact Statement

This paper presents work whose goal is to support research in the field of large language model development. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Ankner, Z., Blakeney, C., Sreenivasan, K., Marion, M., Leavitt, M. L., and Paul, M. Perplexed by perplexity: Perplexity-based data pruning with small reference models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=1GTARJhxtq.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, New York, NY, 2006. ISBN 978-0-387-31073-2.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Coleman, C., Yeh, C., Mussmann, S., Mirzasoleiman, B., Bailis, P., Liang, P., Leskovec, J., and Zaharia, M. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJg2b0VYDr.

Dasgupta, S., Cohn, T., and Baldwin, T. Don't ignore the tail: Decoupling top-k probabilities for efficient language model distillation, 2026. URL https://arxiv.org/abs/2602.20816.

Datta, J., Doll, N., Ramadan, Q., and Boukhers, Z. Exploring the limits of model compression in LLMs: A knowledge distillation study on QA tasks. In Béchet, F., Lefèvre, F., Asher, N., Kim, S., and Merlin, T. (eds.), *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 477–483, Avignon, France, August 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.sigdial-1.39/.

Du, N., Huang, Y., Dai, A. M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A. W., Firat, O., Zoph, B., Fedus, L., Bosma, M. P., Zhou, Z., Wang, T., Wang, E., Webster, K., Pellat, M., Robinson, K., Meier-Hellstern, K., Duke, T., Dixon, L., Zhang, K., Le, Q., Wu, Y., Chen, Z., and Cui, C. GLaM: Efficient scaling of language models with mixture-of-experts. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5547–5569. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/du22c.html.

Elisseeff, A. and Weston, J. A kernel method for multi-labelled classification. In Dietterich, T., Becker, S., and Ghahramani, Z. (eds.), *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/file/39dcaf7a053dc372fbc391d4e6b5d693-Paper.pdf.

Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Gu, Y., Dong, L., Wei, F., and Huang, M. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

Kang, F., Just, H. A., Sun, Y., Jahagirdar, H., Zhang, Y., Du, R., Sahu, A. K., and Jia, R. Get more for less: Principled data selection for warming up fine-tuning in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QmYNBVukex.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Kim, Y. and Rush, A. M. Sequence-level knowledge distillation. In Su, J., Duh, K., and Carreras, X. (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327,

Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Laurençon, H., Saulnier, L., Wang, T., Akiki, C., del Moral, A. V., Scao, T. L., Werra, L. V., Mou, C., Ponferrada, E. G., Nguyen, H., Frohberg, J., Šaško, M., Lhoest, Q., McMillan-Major, A., Dupont, G., Biderman, S., Rogers, A., allal, L. B., Toni, F. D., Pistilli, G., Nguyen, O., Nikpoor, S., Masoud, M., Colombo, P., de la Rosa, J., Villegas, P., Thrush, T., Longpre, S., Nagel, S., Weber, L., Muñoz, M., Zhu, J., Strien, D. V., Alyafeai, Z., Almubarak, K., Vu, M. C., Gonzalez-Dios, I., Soroa, A., Lo, K., Dey, M., Suarez, P. O., Gokaslan, A., Bose, S., Adelani, D., Phan, L., Tran, H., Yu, I., Pai, S., Chim, J., Lepercq, V., Ilic, S., Mitchell, M., Luccioni, S. A., and Jernite, Y. The bigscience roots corpus: A 1.6tb composite multilingual dataset, 2023. URL https://arxiv.org/abs/2303.03915.

Li, M., Zhang, Y., Li, Z., Chen, J., Chen, L., Cheng, N., Wang, J., Zhou, T., and Xiao, J. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning, 2024. URL https://arxiv.org/abs/2308.12032.

Li, Z. and Hoiem, D. Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12):2935–2947, December 2018. ISSN 0162-8828. doi: 10.1109/TPAMI.2017.2773081. URL https://doi.org/10.1109/TPAMI.2017.2773081.

Liu, A. H., Khandelwal, K., Subramanian, S., Jouault, V., Rastogi, A., Sadé, A., Jeffares, A., Jiang, A., Cahill, A., Gavaudan, A., Sablayrolles, A., Héliou, A., You, A., Ehrenberg, A., Lo, A., Eliseev, A., Calvi, A., Soori-yarachchi, A., Bout, B., Rozière, B., Monicault, B. D., Lanfranchi, C., Barreau, C., Courtot, C., Grattarola, D., Dabert, D., de las Casas, D., Chane-Sane, E., Ahmed, F., Berrada, G., Ecrepont, G., Guinet, G., Novikov, G., Kunsch, G., Lample, G., Martin, G., Gupta, G., Ludziejewski, J., Rute, J., Studnia, J., Amar, J., Delas, J., Roberts, J. S., Yadav, K., Chandu, K., Jain, K., Aitchison, L., Fainsin, L., Blier, L., Zhao, L., Martin, L., Saulnier, L., Gao, L., Buyl, M., Jennings, M., Pellat, M., Prins, M., Poirée, M., Guillaumin, M., Dinot, M., Futeral, M., Darrin, M., Augustin, M., Chiquier, M., Schimpf, M., Grinsztajn, N., Gupta, N., Raghuraman, N., Bousquet, O., Duchenne, O., Wang, P., von Platen, P., Jacob, P., Wambergue, P., Kurylowicz, P., Muddireddy, P. R., Chagniot, P., Stock, P., Agrawal, P., Torroba, Q., Sauvestre, R., Soletskyi, R., Menneer, R., Vaze, S., Barry, S., Gandhi, S., Waghjale, S., Gandhi, S., Ghosh, S., Mishra, S., Aithal, S., Antoniak, S., Scao, T. L., Cachet, T., Sorg, T. S., Lavril, T., Saada, T. N., Chabal, T., Foubert, T., Robert, T., Wang, T., Lawson, T., Bewley, T., Bewley, T., Edwards, T., Jamil, U., Tomasini, U., Nemychnikova, V., Phung, V., Maladière, V., Richard, V., Bouaziz, W., Li, W.-D., Marshall, W., Li, X., Yang, X., Ouahidi, Y. E., Wang, Y., Tang, Y., and Ramzi, Z. Ministral 3, 2026. URL https://arxiv.org/abs/2601.08584.

Ma, M., Chochlakis, G., Pandiyan, N. M., Thomason, J., and Narayanan, S. Large language models do multi-label classification differently. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2472–2495, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.126. URL https://aclanthology.org/2025.emnlp-main.126/.

Muennighoff, N., Rush, A. M., Barak, B., Scao, T. L., Tazi, N., Piktus, A., Pyysalo, S., Wolf, T., and Raffel, C. Scaling data-constrained language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=j5BuTrEj35.

Müller, R., Kornblith, S., and Hinton, G. E. When does label smoothing help? In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/f1748d6b0fd9d439f71450117eba2725-Paper.pdf.

OLMo, T., Walsh, P., Soldaini, L., Groeneveld, D., Lo, K., Arora, S., Bhagia, A., Gu, Y., Huang, S., Jordan, M., Lambert, N., Schwenk, D., Tafjord, O., Anderson, T., Atkinson, D., Brahman, F., Clark, C., Dasigi, P., Dziri, N., Ettinger, A., Guerquin, M., Heineman, D., Ivison, H., Koh, P. W., Liu, J., Malik, S., Merrill, W., Miranda, L. J. V., Morrison, J., Murray, T., Nam, C., Poznanski, J., Pyatkin, V., Rangapur, A., Schmitz, M., Skjonsberg, S., Wadden, D., Wilhelm, C., Wilson, M., Zettlemoyer, L., Farhadi, A., Smith, N. A., and Hajishirzi, H. 2 olmo 2 furious, 2025. URL https://arxiv.org/abs/2501.00656.

Pestian, J. P., Brew, C., Matykiewicz, P., Hovermale, D., Johnson, N., Cohen, K. B., and Duch, W. A shared task involving multi-label classification of clinical free text. In Cohen, K. B., Demner-Fushman, D., Friedman, C., Hirschman, L., and Pestian, J. (eds.), *Biological, translational, and clinical language processing*, pp. 97–104, Prague, Czech Republic, June 2007. As-

sociation for Computational Linguistics. URL https://aclanthology.org/W07-1013/.

Raman, M., Mani, P., Liang, D., and Lipton, Z. For distillation, tokens are not all you need. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023. URL https://openreview.net/forum?id=2fc5GOPYip.

Rawat, A. S., Sadhanala, V., Rostamizadeh, A., Chakrabarti, A., Jitkrittum, W., Feinberg, V., Kim, S., Harutyunyan, H., Saunshi, N., Nado, Z., Shivanna, R., Reddi, S. J., Menon, A. K., Anil, R., and Kumar, S. A little help goes a long way: Efficient LLM training by leveraging small LMs, 2026. URL https://openreview.net/forum?id=UrGsJphPnJ.

Schapire, R. E. and Singer, Y. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2):135–168, May 2000.

Soboleva, Daria, e. a. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. https://www.cerebras.net/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama, 2023.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision, 2015. URL https://arxiv.org/abs/1512.00567.

Team, Q. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Tsoumakas, G. and Katakis, I. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3:1–13, 09 2009. doi: 10.4018/jdwm.2007070101.

Wang, Q. and Zhou, J. Topkd: Top-scaled knowledge distillation. *ArXiv*, abs/2508.04539, 2025. URL https://api.semanticscholar.org/CorpusID:280536469.

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., and Grave, E. Ccnet: Extracting high quality monolingual datasets from web crawl data, 2019. URL https://arxiv.org/abs/1911.00359.

Zhang, Q., Guo, Y., and Xiang, Y. Continual distillation learning: Knowledge distillation in prompt-based continual learning, 2025a. URL https://arxiv.org/abs/2407.13911.

Zhang, X., Liangyu, X., Duan, F., Zhou, Y., Wang, S., Weng, R., Wang, J., and Cai, X. Preference curriculum: LLMs should always be pretrained on their preferred data. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21181–21198, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1091. URL https://aclanthology.org/2025.findings-acl.1091/.