Bonsai: A Small Ternary-Weight Language Model

Hespere AI

Abstract

We introduce Bonsai, a 0.5 billion parameter ternary-weight language model. Through innovative quantization-aware training methods and careful data curation, we pretrain Bonsai in under 5 billion total tokens. Through our work, Bonsai attains competitive benchmark results among small language models such as Qwen 2.5 0.5B, and outperforms all other open ternary-weight models in its parameter class. Our results show that Bonsai is one of the first ternary weight models to match leading full precision models of similar parameter count. Our results highlight the increasing feasibility of extreme low bit quantized models with minimal compute. Bonsai is available at https://huggingface.co/hespere-ai/Bonsai.

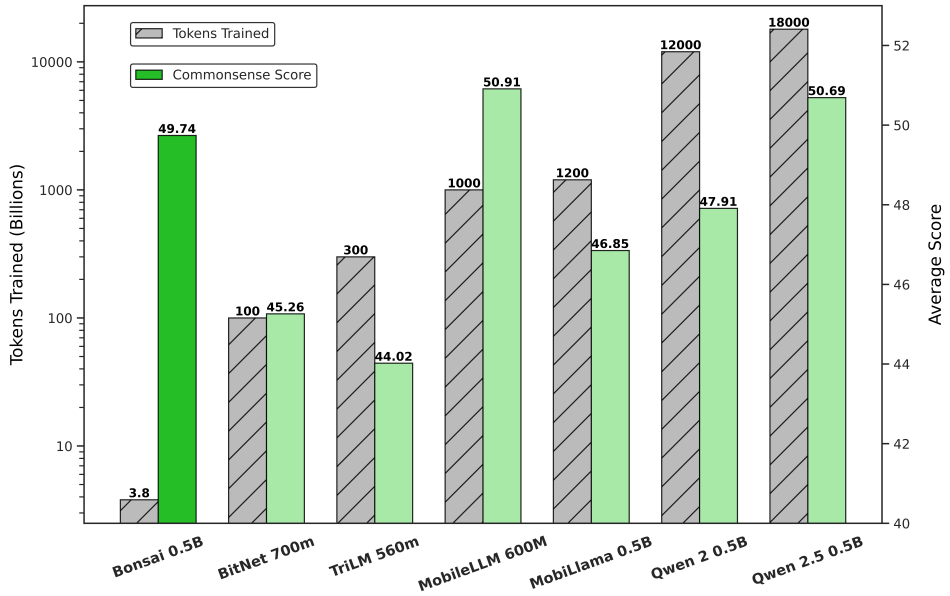

Figure 1: Model performance on a suite of 6 benchmarks compared to tokens trained.

1 Introduction

Large language models (LLMs) have become ubiquitous, with applications ranging from chatbots for everyday use, to cutting-edge scientific research, and more. However, most language models are computationally intensive to operate, leading to high costs and concerns about environmental impact. In response to concerns about computational cost, recent research [Ma et al., 2024] has shown that extremely low-bit (ternary-weight) language models offer a potential solution, operating with the same performance as full-precision models while being much more efficient. Yet, due to the limited expressive range of ternary weights, [Ma et al., 2024] find that post-training quantization is

suboptimal for ternary-weight models, meaning that an expensive and compute-intensive pre-training process is required.

In this paper, we aim to show that the pretraining process need not be expensive and compute-intensive by introducing Bonsai, an efficiently trained and performant ternary-weight model. Bonsai is based on the Llama architecture [Grattafiori et al., 2024] [Touvron et al., 2023], following the configuration specified by Danube 3 [Pfeiffer et al., 2024]. It is one of the first ternary-weight models to match fully-trained full-precision models at the same parameter count. This performance and efficiency is achieved utilizing a two-stage quantization-aware training pipeline, which allows for the model to be trained in under 5 billion tokens, a fraction of the compute compared to other methods. Our method sets a new standard for resource efficient quantization-aware training.

Our contributions are as follows:

1. We release a competitive ternary small language model, allowing for further development of the ternary ecosystem. Our release also allows researchers to further explore the behavior of extreme low-bit quantized language models.
2. We advance the study of ternary quantization via a novel quantization-aware training scheme for ternary models.
3. We show that competitive ternary models can be achieved under extremely low compute requirements.

Model	**Precision**	**Total Bits**
Bonsai	1.58b	$\sim 2.23 \times 10^9$ (1x)
TriLM 560M	1.58b	$\sim 2.76 \times 10^9$ (1.23x)
BitNet 700M	1.58b	$\sim 2.56 \times 10^9$ (1.15x)
Qwen 2.5 0.5B	fp16	$\sim 7.90 \times 10^9$ (3.54x)
Qwen 2 0.5B	fp16	$\sim 7.90 \times 10^9$ (3.54x)
MobiLlama 0.5B	fp16	$\sim 8.56 \times 10^9$ (3.83x)
Falcon 3 1B	1.58b	$\sim 9.44 \times 10^9$ (4.23x)
MobileLLM 600M	fp16	$\sim 9.65 \times 10^9$ (4.32x)

Table 1: Model size comparison in term of estimated total number of bits per model.

2 Quantization-Aware Training

This section provides a detailed overview of the quantization-aware training process for Bonsai. We note that unlike [Ma et al., 2024], we do not include any additional norms in the model structure. We also do not attempt to quantize activations; all activations are left in 16-bit precision.

2.1 Stage 1: Training

For quantization-aware training, we begin by decomposing our weights into two components: a channel-wise scaling vector $v \in \mathbb{R}^m$, and a ternary matrix T initialized as

$$T = \text{sign}(\text{TWN}(W)) \in \{-1, 0, 1\}^{m \times n}$$

where TWN denotes the Ternary Weight Networks quantizer [Li et al., 2022]. The scaling vector v is initialized such that each ith element v_i corresponds to the 2-norm of the ith row of W; $v_i = ||w^i_{fp}||$. In the interest of stability, the forward pass of the linear layer is modified to normalize the weight matrix row-wise before multiplying by the scaling factor, in order to keep the magnitudes of each weight matrix consistent with respect to the scaling vector (let the normalization operation be denoted

as $\|\cdot\|_{row}$).The T matrix and v vector are made trainable with T using a straight-through estimator; the modified linear layer's forward pass is thus

$$y = (v \odot \|\mathrm{Q}(T))\|_{row})x + b \tag{1}$$

where $\odot$ denotes the element-wise product and Q denotes the following quantizer

$$\mathrm{Q}(T) = \lfloor T.\mathrm{clamp}(-1, 1) \rceil \tag{2}$$

We note that post-quantization, the normalization operation can be removed from the layer, with the normalization factors being absorbed into the row-wise scales.

We train the quantized model end-to-end using cross-entropy loss. With respect to training data, we utilize a mix of data filtered from FineWeb-Edu [Lozhkov et al., 2024] to have an educational score greater than 4, and DCLM-Pro [Zhou et al., 2024]. We pack the curated data into sequences of length 2048. The weights are initially trained for 27,000 steps with a global batch size of 32 purely on DCLM-Pro, which amounts to a little more than 1.5 billion tokens. We then introduce the filtered data from Fineweb-Edu into our data mixture to make a roughly 60% DCLM-Pro and 40% filtered Fineweb-Edu mixture before continuing training for 16,000 steps with a larger global batch size of 64. Training is performed with the cautious AdamW optimizer [Liang et al., 2025], with a learning rate of 0.01 and a cosine learning rate schedule with linear warm-up for both phases. In total, we estimate that we train for around 3.8 billion tokens.

2.2 Stage 2: Scale and Norm Tuning

Following the observations of [Li et al., 2023], [Zeng et al., 2024] and [Chen et al., 2024] on the efficacy of further tuning of quantization parameters, we utilize a second stage of quantization in which we train the norms and scales of our model, while we freeze the rest of the model. Adjustments to the scales and norms allow the distribution of the model shift slightly, which improves performance while minimizing forgetting. We utilize 128 samples from DCLM [Li et al., 2024] to perform this step, and optimize for 10 epochs with a global batch size of 64. We again use the cautious AdamW optimizer with a learning rate of 0.01, and a cosine learning rate schedule with linear warm-up.

3 Evaluations

We compare Bonsai to baseline LLMs in its parameter range to gauge its performance. The baselines include the Qwen 2.5 [Qwen et al., 2025] and Qwen 2 [Yang et al., 2024] series, MobiLlama [Thawakar et al., 2024], MobileLLM [Liu et al., 2024], Falcon 3 [Team, 2024], Spectra [Kaushal et al., 2024] and Bitnet [Ma et al., 2024]. Our baselines include high-performing full-precision models, along with ternary baselines.

We evaluate Bonsai on various common sense reasoning benchmarks using the Language Model Evaluation Harness [Gao et al., 2021], which includes: ARC-Challenge [Clark et al., 2018], ARC-Easy [Clark et al., 2018], HellaSwag [Zellers et al., 2019], PiQA [Bisk et al., 2019], OpenBookQA [Mihaylov et al., 2018], and Winogrande [Sakaguchi et al., 2021]. We also evaluate MMLU [Hendrycks et al., 2021] using the formulation provided by lighteval [Fourrier et al., 2023] due to the small size of the model. All models were evaluated in the same environment using normalized accuracy metrics for all benchmarks outside of Winogrande, which uses raw accuracy.

3.1 Full-Precision Comparisons

Table 2 presents the zero-shot evaluation results. In this evaluation, Bonsai demonstrates competitive performance, although it lags behind the SOTA, Qwen 2.5 by around $\sim 1.5\%$. We order models via their average scores across the 7 benchmarks. Bonsai performs on par or better than baseline models such as MobiLlama 0.5B and Qwen 2 0.5B on average. However, we note that there are areas where Bonsai observes more significant weakness, such as MMLU, where Bonsai places the lowest in the group, thus suggesting that there is room for improvement in knowledge-based benchmarks.

3.2 Ternary Comparisons

We additionally include a comparison to the current best ternary weight models that we were able to evaluate. Since certain models were unable to be evaluated on MMLU, we restrict our

Model	ARC-c	ARC-e	HS.	OBQA	PiQA	Wino.	MMLU	Avg
MobiLlama 0.5B	26.62	46.68	51.66	30.00	71.65	54.50	28.61	44.25
Qwen 2 0.5B	28.84	50.29	49.12	33.00	69.26	56.99	31.78	45.61
MobileLLM 600M	29.01	56.65	55.35	34.00	71.65	59.75	31.40	48.13
Qwen 2.5 0.5B	32.25	58.29	52.18	35.40	69.91	56.12	33.40	48.22
Bonsai	33.36	57.95	48.04	34.00	70.24	54.85	30.28	46.96

Table 2: Zero-shot performance on standard benchmarks versus full-precision baselines.

Model	ARC-c	ARC-e	HS.	OBQA	PiQA	Wino.	Avg
Bitnet 700M	25.34	46.25	43.8	32.40	68.23	55.56	45.26
TriLM 560M	25.68	45.54	41.53	30.60	67.25	53.51	44.02
Falcon 3 1B Instruct 1.58bit	28.50	51.94	44.46	32.00	66.87	54.06	46.03
Bonsai	33.36	57.95	48.04	34.00	70.24	54.85	49.74

Table 3: Zero-shot performance on restricted set of standard benchmarks versus ternary models.

benchmarks to six commonsense reasoning benchmarks: ARC-Challenge, ARC-Easy, HellaSwag, PiQA, OpenbookQA, and Winogrande. Each model was evaluated with the Language Model Evaluation Harness [Gao et al., 2021].

Cost	Bonsai (500M)	GPT2-774M (Karpathy)
Tokens Trained (Billions)	3.8	150
Estimated H100 Hours	40	1,080
USD	$70	$2000
Avg Performance	49.74	50.71

Table 4: Comparison of training costs and performance between Bonsai and GPT2 774M (Karpathy reproduction on Fineweb). Performance evaluations were performed using Language Model Evaluation Harness, using the restricted set of ARC-Challenge, ARC-Easy, HellaSwag, PiQA, OpenbookQA, and Winogrande.

4 Conclusions

This paper introduced Bonsai, a 500-million-parameter ternary weight model which demonstrates a new frontier in small model performance. By leveraging a multistage quantization-aware-training pipeline, Bonsai showed significant improvements in standard benchmarks compared to previously released ternary models, competing with full precision counterparts. Through this paper, we hope to help advance the study and adoption of ternary weight networks.

References

Y. Bisk, R. Zellers, R. L. Bras, J. Gao, and Y. Choi. Piqa: Reasoning about physical commonsense in natural language. In *AAAI Conference on Artificial Intelligence*, 2019. URL https://api.semanticscholar.org/CorpusID:208290939.

M. Chen, W. Shao, P. Xu, J. Wang, P. Gao, K. Zhang, and P. Luo. Efficientqat: Efficient quantization-aware training for large language models, 2024. URL https://arxiv.org/abs/2407.11062.

P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018.

C. Fourrier, N. Habib, H. Kydlíček, T. Wolf, and L. Tunstall. Lighteval: A lightweight framework for llm evaluation, 2023. URL https://github.com/huggingface/lighteval.

L. Gao, J. Tow, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, K. McDonell, N. Muennighoff, J. Phang, L. Reynolds, E. Tang, A. Thite, B. Wang, K. Wang, and A. Zou. A framework for few-shot language model evaluation, Sept. 2021. URL https://doi.org/10.5281/zenodo.5371628.

A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Wyatt, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Guzmán, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Thattai, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. Kloumann, I. Misra, I. Evtimov, J. Zhang, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Prasad, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, K. Lakhotia, L. Rantala-Yeary, L. van der Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo, L. Blecher, L. Landzaat, L. de Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kardas, M. Tsimpoukelli, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M. K. Singh, M. Hassan, N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, N. Zhang, O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Krishnan, P. S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R. S. Cabral, R. Stojnic, R. Raileanu, R. Maheswari, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly, R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S. S. Kim, S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhende, S. Batra, S. Whitman, S. Sootla, S. Collot, S. Gururangan, S. Borodinsky, T. Herman, T. Fowler, T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami, V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet, V. Do, V. Vogeti, V. Albiero, V. Petrovic, W. Chu, W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X. Wang, X. E. Tan, X. Xia, X. Xie, X. Jia, X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z. D. Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Srivastava, A. Jain, A. Kelsey, A. Shajnfeld, A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenberg, A. Baevski, A. Feinstein, A. Kallet, A. Sangani, A. Teo, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho, A. Poulton, A. Ryan, A. Ramchandani, A. Dong, A. Franco, A. Goyal, A. Saraf, A. Chowdhury, A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang, B. Loyd, B. D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence, B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Liu, C. Wang, C. Kim, C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feichtenhofer, C. Gao, D. Civin, D. Beaty, D. Kreymer, D. Li, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss, D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood, E.-T. Le, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Kokkinos, F. Ozgenel, F. Caggioni, F. Kanayet, F. Seide, G. M. Florez, G. Schwarz, G. Badeer, G. Swee, G. Halpern, G. Herman, G. Sizov, Guangyi, Zhang, G. Lakshminarayanan, H. Inan, H. Shojanazeri, H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspegren, H. Goldman, H. Zhan, I. Damlaj, I. Molybog, I. Tufanov, I. Leontiadis, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski, J. Kohli, J. Lam, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein, J. Teboul, J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J. Ginsburg, J. Wang, K. Wu, K. H. U, K. Saxena, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan, K. Michelena, K. Li, K. Jagadeesh, K. Huang, K. Chawla, K. Huang, L. Chen, L. Garg, L. A, L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani, M. Bhatt, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi, M. Keneally, M. Liu, M. L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M. J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N. Mehta, N. P. Laptev, N. Dong, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Parthasarathy, R. Li, R. Hogan, R. Battey, R. Wang, R. Howes, R. Rinott, S. Mehta, S. Siby, S. J. Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Mahajan, S. Verma, S. Yamamoto, S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S. C. Zha, S. Patil, S. Shankar, S. Zhang, S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S. Govindaprasad, S. Gupta, S. Deng, S. Cho, S. Virk, S. Subramanian,

S. Choudhury, S. Goldman, T. Remez, T. Glaser, T. Best, T. Koehler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou, T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V. S. Kumar, V. Mangla, V. Ionescu, V. Poenaru, V. T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable, X. Tang, X. Wu, X. Wang, X. Wu, X. Gao, Y. Kleinman, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li, Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y. Zhao, Y. Hao, Y. Qian, Y. Li, Y. He, Z. Rait, Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, Z. Zhao, and Z. Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

A. Kaushal, T. Vaidhya, T. Pandey, A. Bhagat, and I. Rish. TriLM vs floatLM: Ternary LLMs are more performant than quantized FP16 LLMs. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024. URL https://openreview.net/forum?id=gvaDL9omKU.

F. Li, B. Liu, X. Wang, B. Zhang, and J. Yan. Ternary weight networks, 2022. URL https://arxiv.org/abs/1605.04711.

J. Li, A. Fang, G. Smyrnis, M. Ivgi, M. Jordan, S. Gadre, H. Bansal, E. Guha, S. Keh, K. Arora, S. Garg, R. Xin, N. Muennighoff, R. Heckel, J. Mercat, M. Chen, S. Gururangan, M. Wortsman, A. Albalak, Y. Bitton, M. Nezhurina, A. Abbas, C.-Y. Hsieh, D. Ghosh, J. Gardner, M. Kilian, H. Zhang, R. Shao, S. Pratt, S. Sanyal, G. Ilharco, G. Daras, K. Marathe, A. Gokaslan, J. Zhang, K. Chandu, T. Nguyen, I. Vasiljevic, S. Kakade, S. Song, S. Sanghavi, F. Faghri, S. Oh, L. Zettlemoyer, K. Lo, A. El-Nouby, H. Pouransari, A. Toshev, S. Wang, D. Groeneveld, L. Soldaini, P. W. Koh, J. Jitsev, T. Kollar, A. G. Dimakis, Y. Carmon, A. Dave, L. Schmidt, and V. Shankar. Datacomp-lm: In search of the next generation of training sets for language models, 2024. URL https://arxiv.org/abs/2406.11794.

L. Li, Q. Li, B. Zhang, and X. Chu. Norm tweaking: High-performance low-bit quantization of large language models, 2023. URL https://arxiv.org/abs/2309.02784.

K. Liang, L. Chen, B. Liu, and Q. Liu. Cautious optimizers: Improving training with one line of code, 2025. URL https://arxiv.org/abs/2411.16085.

Z. Liu, C. Zhao, F. Iandola, C. Lai, Y. Tian, I. Fedorov, Y. Xiong, E. Chang, Y. Shi, R. Krishnamoorthi, L. Lai, and V. Chandra. Mobilellm: Optimizing sub-billion parameter language models for on-device use cases, 2024. URL https://arxiv.org/abs/2402.14905.

A. Lozhkov, L. Ben Allal, L. von Werra, and T. Wolf. Fineweb-edu: the finest collection of educational content, 2024. URL https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu.

S. Ma, H. Wang, L. Ma, L. Wang, W. Wang, S. Huang, L. Dong, R. Wang, J. Xue, and F. Wei. The era of 1-bit llms: All large language models are in 1.58 bits, 2024. URL https://arxiv.org/abs/2402.17764.

T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.

P. Pfeiffer, P. Singer, Y. Babakhin, G. Fodor, N. Dhankhar, and S. S. Ambati. H2o-danube3 technical report, 2024. URL https://arxiv.org/abs/2407.09276.

Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9): 99–106, 2021.

F.-L. Team. The falcon 3 family of open models, December 2024. URL https://huggingface.co/blog/falcon3.

O. Thawakar, A. Vayani, S. Khan, H. Cholakal, R. M. Anwer, M. Felsberg, T. Baldwin, E. P. Xing, and F. S. Khan. Mobillama: Towards accurate and lightweight fully transparent gpt, 2024.

H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Yang, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, X. Liu, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, Z. Guo, and Z. Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. Hellaswag: Can a machine really finish your sentence?, 2019.

C. Zeng, S. Liu, Y. Xie, H. Liu, X. Wang, M. Wei, S. Yang, F. Chen, and X. Mei. Abq-llm: Arbitrary-bit quantized inference acceleration for large language models, 2024. URL https://arxiv.org/abs/2408.08554.

F. Zhou, Z. Wang, Q. Liu, J. Li, and P. Liu. Programming every example: Lifting pre-training data quality like experts at scale. *arXiv preprint arXiv:2409.17115*, 2024.