Towards Efficient Pre-training: Exploring FP4 Precision in Large Language Models

Jiecheng Zhou [1 2] **Ding Tang** [2] **Rong Fu** [2] **Boni Hu** [2] **Haoran Xu** [2] **Yi Wang** [2] **Zhongling Su** [2] **Liang Liu** [2] **Zhilin Pei** [2] **Hengjie Li** [2] **Xingcheng Zhang** [2] **Weiming Zhang** [1]

Abstract

The burgeoning computational demands for training large language models (LLMs) necessitate efficient methods, including quantized training, which leverages low-bit arithmetic operations to reduce costs. While FP8 precision has shown potential, leveraging FP4 remains challenging due to inherent quantization errors and limited representation capability. Based on the Transformer architecture, we present an FP4 training scheme for LLMs, overcoming these obstacles through mixed-precision quantization strategies tailored for different modules and training stages. This allows us to apply the precision level suitable to distinct components within the model, ensuring that multi-head attention and linear layers are handled appropriately. Our pretraining recipe ensures stability in backpropagation by incorporating fine-grained quantization methods with a target precision training schedule. Experimental results demonstrate that our FP4 training scheme achieves accuracy comparable to BF16 and FP8, with smaller theoretical computational cost. With the advent of next-generation hardware supporting FP4, our method sets the foundation for efficient ultra-low precision training.

1. Introduction

Recent advancements in large language models, including GPT (Radford, 2018; Floridi & Chiriatti, 2020; Achiam et al., 2023), DeepSeek (Liu et al., 2024), Llama (Touvron et al., 2023) and OPT (Zhang et al., 2023), have demonstrated strong generalization capabilities across various tasks (Stiennon et al., 2020; Alexey, 2020). Among these advancements, pretraining on large-scale unlabeled data has proven to be critical for ensuring model performance (Radford, 2018; Dong et al., 2019). Increasing the model size and the dataset scale can enhance performance (Kaplan et al., 2020; Hoffmann et al., 2022), but this improvement comes with significant computational costs. To address this, numerous methods have been proposed to accelerate the pretraining process (Duan et al., 2024).Particularly, low-precision computation serves as an efficient acceleration technique. This approach quantizes the inputs of computationally intensive operators to a specified low-bit width, leveraging low-bit arithmetic units to speed up training.

Previous research on low-precision training has primarily focused on deep learning models. However, these methods do not fully consider the characteristics of large language model pre-training, which has unique training methods and model architectures. The good news is that next-generation hardware will support FP4 and FP8 format (Nvidia). Studies like (Peng et al., 2023; Micikevicius et al., 2022b; NVIDIA; Fishman et al., 2024; Xi et al.) have demonstrated the capability of 8-bit computation for LLM pretraining. However, the application of FP4 tensor cores in LLM pretraining remains unexplored. Compared to INT4, FP4 offers a larger numerical representation space, making it possible to further reduce the bit width in large-scale model pretraining. However, the limited number of bits in FP4 format introduces significant quantization errors, making the application of FP4 to pretraining highly challenging.

From the perspective of LLM structure, it has been observed that different modules and computational components exhibit varying levels of sensitivity. Given the critical role of the multi-head attention (MHA) mechanism in the Transformer architecture, it is imperative to implement specific strategies to ensure the accuracy of attention modules. As the volume of data continues to grow, the gradient values tend to decrease, making FP4 quantization more prone to underflow, thus hindering parameter updates. Based on these observations, we propose a novel FP4 mixed-precision large language model pretraining recipe. Specifically, we lever-

[*]Equal contribution [1]School of Information Science and Technology, University of Science and Technology of China, Anhui, China [2]Shanghai Artificial Intelligence Laboratory, Shanghai, China. Correspondence to: Weiming Zhang <zhangwm@ustc.edu.cn>, Rong Fu <furong@pjlab.org.cn>, Jiecheng Zhou <zhoujiecheng@ustc.edu.cn>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

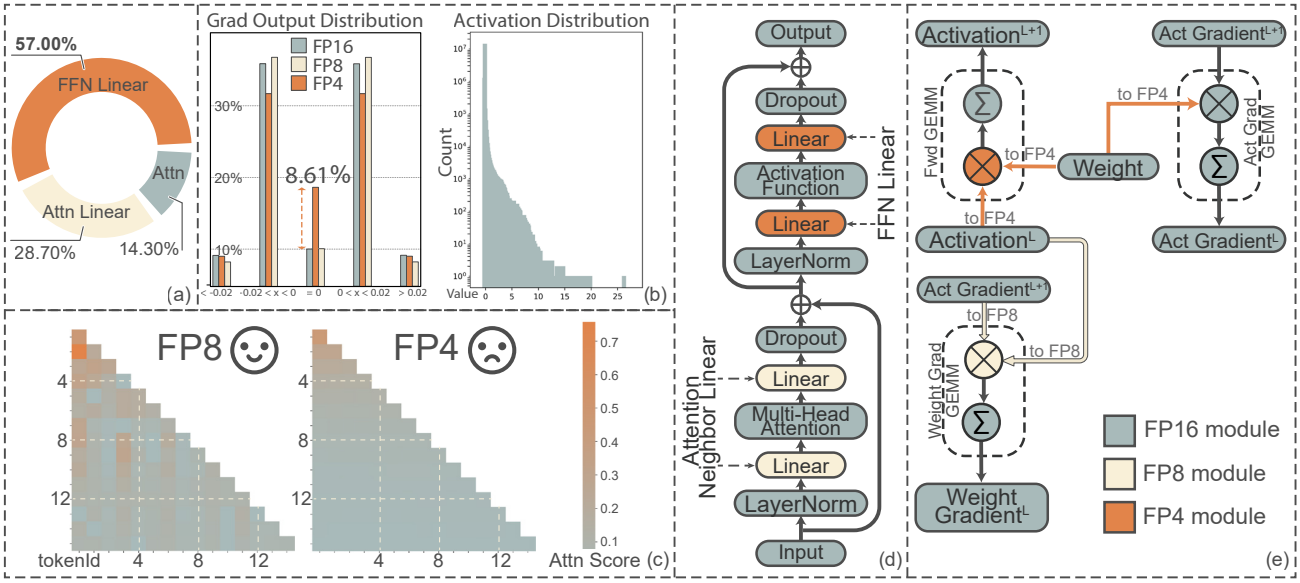

Figure 1. (a) shows the proportion of computational overhead for the main computation components of a transformer block when using the LLaMA 7B configuration with a sequence length of 4K. (b) shows the distribution of activations and gradients after the GPT-large model has been trained to approximately 10B tokens. (c) shows the heatmap of attention scores when using different training strategies. (d) and (e) illustrate our training scheme, which will be detailed in Section 3.

age a per-block quantization strategy and employ different quantization approaches across modules and training stages to enable FP4 model training. The approaches enable better exploitation of the computational improvements brought by future hardware advancements.

In this paper, we explored the use of FP4 precision in language model pretraining and, for the very first time, proposed an effective mixed precious pretraining strategy. First, considering the distinct requirements of different modules, we applied tailored quantization strategies to preserve the precision of MHA execution. Second, as backpropagation has shown high sensitivity to precision, we employed finer-grained quantization methods to ensure accurate parameter updates during the backward pass. Finally, we adopted a 2-stage target precision training schedule to eliminate the impact of quantization noise on the model.

2. Related Work

Low-precision training enhances deep learning efficiency by reducing computational costs. Many existing studies focus on the training of deep neural networks (DNNs) (Wang et al., 2018; Chmiel et al., 2023; Sun et al., 2019; Xi et al., 2023; Fu et al., 2021), whose architecture and performance differ from LLM pre-training. In the context of low-precision training for large model pre-training, some progress has been made in FP8. For example, (Micikevicius et al., 2022b) introduced new FP8 floating-point formats (E4M3 and E5M2), and (Fishman et al., 2024) extends FP8 to trillion-token large-scale model pretraining. In terms of FP4 training, (Wang et al., 2025) improved FP4 computational precision using a differentiable quantization estimator and outlier clamping and compensation strategy. However, most existing methods fail to fully account for the varying sensitivity to precision across different model modules.

3. Methods

Our objective is to maximize the efficiency of low-precision computations based on the characteristics of LLMs. As shown in Fig.1(a), the computational cost of three key components in a transformer model is analyzed, with FFN accounting for 57%. Considering both the computational cost and impact on performance, we meticulously design three corresponding training schemes: 3.1 Attention-protected Neighbor Linear, 3.2 Gradient-sensitive Linear, and 3.3 Target Precious Training Schedule. These schemes fully leverage hardware acceleration while keeping precision loss within an acceptable range during training.

3.1. Attention-protected Neighbor Linear

As the core component of the Transformer model, the attention mechanism is highly sensitive to precision. Quantization errors introduced by low-precision training can accu-

mulate over time, eventually disrupting the function of the attention mechanism. As shown in Fig. 1(c), an undisturbed attention mechanism identifies tokens 0, 3, 6, and 9 as more important. However, under FP4 training, the result becomes nearly uniformly distributed, preventing the model from distinguishing which tokens are significant. This makes it difficult for the model to differentiate the importance of tokens, thereby affecting the convergence speed.

To ensure the proper functioning of the attention mechanism and enable the model to correctly evaluate the importance of each token, we employ FP8 precision for the computation of QKV and the output projection to "protect" the accurate execution of the attention mechanism, as shown in Fig. 1(d).

3.2. Gradient-sensitive FFN Linear

Weight gradient computation is more sensitive to errors compared to forward computation, due to the fact that both gradients and activations contribute to the error. For gradients, since many values are around 0.02, especially as training progresses and gradient magnitudes decrease, underflow is likely to occur. As can be seen in the left of Fig.1(b), there is an 8.6% difference between FP4 and FP8/FP16, thus requiring a more accurate representation. For activations, we observe that underflow occurs approximately 18% of the time between FP4 and FP8/FP16. This is largely due to the relatively large range of values, as can be seen in the right of Fig. 1(b). Therefore, a more accurate representation is also needed. Additionally, optimizers use the gradients to update model parameters. Based on the above discussion, for the weight gradient computation of model weights, we adopt FP8 precious computation, as shown in the bottom left corner of Fig.1(e).

Furthermore, for the activation gradient computation (the top right corner of Fig.1(e)), we find that quantizing gradients significantly impacts the convergence of model training. There is always a nonlinear operation between the linear layers, which requires more precise numerical representations. Furthermore, quantization errors accumulate iteratively through the chain rule during backpropagation, ultimately hindering the convergence of model training.

Lastly, in our experiments, we observed that quantization noise increases as the model size and the amount of data grow (a detailed explanation can be found in Appendix B). This occurs because, when the model reaches a certain level of accuracy, coarse-grained low-precision tensors can no longer currently represent the parameter space and input information. Therefore, we adopt a more conservative quantization approach to maintain stable training in forward computation. As shown in the top left corner of Fig.1(e). To ensure efficient hardware implementation, we use per-block quantization strategies where the block size is set to 128.

3.3. Target Precious Training Schedule

When using low-precision training throughout the entire process, there tends to be a performance gap between the low-precision model and the FP16 model, as shown in Fig.2. The validation loss curves exhibit a parallel trend. Although the gap between the two curves is very small, the difference in downstream tasks, such as wikitext perplexity (PPL), can be more pronounced, reaching up to approximately 6.3 compared to the model trained with FP16. This is likely due to compromises the model makes to adapt to the noise introduced by quantization during low-precision training. To address this issue, we employ a Target Precious Training Schedule, which involves two stages: continuing the FP4 pretraining process with FP16 for a short period. This accounts for only 5% to 10% of the total training steps, allowing the model to return to an ideal state.

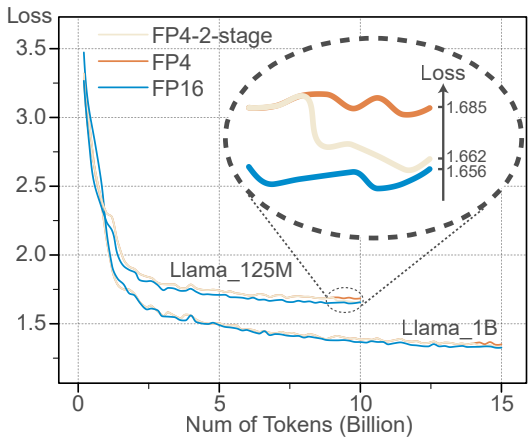

Figure 2. Loss curve for the Target Precious Training Schedule.

4. Experiment

In this section, we evaluate the proposed FP4 training method across language models of various sizes. The detailed model training configurations and hyperparameter settings are provided in Appendix B. Section 4.1 presents the main results, showcasing the model's performance on downstream tasks. Section 4.2 presents the ablation study to demonstrate the effectiveness of our training method.

4.1. Main Result

We validate the proposed FP4 pretraining method on two large language models, using the widely adopted GPT-2 and LLaMA architectures. The GPT-2 and LLaMA models are pretrained on the RedPajama-WikiText (Weber et al., 2025) dataset within the Megatron framework and evaluate their text generation capabilities on wikiText(Merity et al., 2016). Additionally, we assess their natural language understanding

Table 1. Comparison of FP4 and FP16 Training Results

Model	Method	Val Loss	Val PPL	Text Gen	Natural Language Understanding (GLUE)								
				WikiText	cola	sst2	mrpc	stsb	rte	wnli	qnli	mnli	qqp
2*GPT2 125M	Ours	1.706	5.507	50.98	0.2663	0.8704	0.7549 / 0.8322 / 0.7936	0.7681 / 0.7658	0.5704	0.3380	0.8473	0.7554 / 0.7705	0.8777 / 0.8403
	FP16-baseline	1.705	5.503	50.14	0.2290	0.8796	0.7647 / 0.8395 / 0.8021	0.7798 / 0.7808	0.5884	0.3099	0.8548	0.7613 / 0.7695	0.8799 / 0.8437
2*GPT2 335M	Ours	1.549	4.705	37.62	0.2565	0.8899	0.7647 / 0.8362 / 0.8004	0.8159 / 0.8122	0.6029	0.2535	0.8611	0.7798 / 0.7874	0.892 / 0.8572
	FP16-baseline	1.556	4.739	38.39	0.3002	0.8819	0.7745 / 0.8419 / 0.8082	0.8266 / 0.8298	0.6209	0.1831	0.8726	0.7799 / 0.7889	0.8929 / 0.8508
2*GPT2 774M	Ours	1.431	4.181	30.01	0.3473	0.9002	0.7745 / 0.8472 / 0.8108	0.8305 / 0.8336	0.6498	0.2676	0.8742	0.7995 / 0.808	0.8984 / 0.8656
	FP16-baseline	1.430	4.178	28.36	0.3708	0.8922	0.7794 / 0.8454 / 0.8124	0.8347 / 0.8353	0.6498	0.2254	0.8911	0.8078 / 0.813	0.9012 / 0.8683

abilities on the GLUE(Wang, 2018) benchmark.

We train approximately 10B tokens on GPT-2-small and GPT-2-mid and around 25B tokens on GPT-2-large. The final validation loss and validation perplexity (PPL) are presented in Table 1, showing that the pretraining results obtained with our method exhibit almost no performance difference compared to models trained using FP16. In addition to training loss, the downstream task performance of the same pretrained models demonstrates that the average accuracy of FP4-trained models is comparable to that of FP16-trained models.

4.2. Ablation Study

We aim to investigate the effect of the module-wise pretraining method introduced in Section 3. For this ablation study, we train the LLaMA2-125M model on approximately 5B tokens. The results in the table indicate that different modules exhibit varying levels of robustness to low precision. Additionally, we compute the theoretical computation cost for these methods and observe that our approach achieves a lower theoretical computation cost (see Appendix B for details) while maintaining higher performance.

Attention Linear	FFN Linear	FP4 Linear' Backward	Training loss	Val loss	Val ppl	Computation cost
FP4	FP4	FP4	2.2659	1.7828	5.9467	57.1%
FP4	FP8	FP8	2.2211	1.7543	5.7798	69.6%
FP8	FP4	FP4	2.2562	1.7549	5.7831	60.7%
FP8	**FP4**	**FP8**	**2.2225**	**1.7415**	**5.7062**	**66.1%**
FP16	FP16	FP16	2.1998	1.7097	**5.5273**	100%

Table 2. Ablation studies about different precious on different modules

2*	2*Attention	2*FFN	2*FFN Backward	2*Target Precious	2*Val loss	2*Val ppl	2*Computation Cost
3***Llama 1B**	FP8	FP4	FP8	no	1.3505	3.8596	67.5%
	FP8	FP4	FP8	yes	1.3311	**3.7855**	69.7%
	FP16	FP16	FP16	-	1.3296	**3.7797**	100%
3***Llama 125M**	FP8	FP4	FP8	no	1.6851	5.3933	68.2%
	FP8	FP4	FP8	yes	1.6622	**5.2670**	71.4%
	FP16	FP16	FP16	-	1.6567	**5.2424**	100%

Table 3. Ablation studies about target precious training schedule

Furthermore, to demonstrate the effectiveness of the 2-stage training schedule, we conducted the following ablation experiment, using the same experimental setup as in Section 4.1. The results in the table highlight the importance of the 2-stage training schedule for large-scale model pretraining.

5. Conclusion

We propose an FP4 pre-training scheme for modern large language models. Based on the sensitivity analysis of computational modules and training stages, we adopt a tailored training recipe according to the position of linear modules, together with a Target Precious Training schedule, to ensure stable convergence. Experimental results show that FP4-based training achieves comparable validation loss and downstream task accuracy to traditional FP16 training while reducing computational costs by 30%. Additionally, our ablation studies confirm the importance of adaptive quantization strategies across different model modules and training stages, providing new insights for the further development of low-precision training techniques and efficient training of large language models on next-generation hardware.

6. Limitation

First, due to computational resource limitations, our method has not been validated on larger models and larger datasets to demonstrate its effectiveness. Investigating such scalability remains a critical direction for future research. In addition, since the model adopts a simulated FP4 approach, it is unable to obtain an accurate increase in training efficiency. Lastly, for the sensitive computational components, we employed a simple strategy to ensure numerical precision. In future work, we will explore more customized approaches to enable a broader range of computations to utilize FP4.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alexey, D. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020.

Bengio, Y., Léonard, N., and Courville, A. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Chmiel, B., Banner, R., Hoffer, E., Ben-Yaacov, H., and Soudry, D. Accurate neural training with 4-bit matrix multiplications at standard formats. In *The Eleventh International Conference on Learning Representations*, 2023.

Dao, T., Fu, D., Ermon, S., Rudra, A., and Ré, C. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., and Hon, H.-W. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32, 2019.

Duan, J., Zhang, S., Wang, Z., Jiang, L., Qu, W., Hu, Q., Wang, G., Weng, Q., Yan, H., Zhang, X., et al. Efficient training of large language models on distributed infrastructures: A survey. *arXiv preprint arXiv:2407.20018*, 2024.

Fishman, M., Chmiel, B., Banner, R., and Soudry, D. Scaling fp8 training to trillion-token llms. *arXiv preprint arXiv:2409.12517*, 2024.

Floridi, L. and Chiriatti, M. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

Fu, Y., Guo, H., Li, M., Yang, X., Ding, Y., Chandra, V., and Lin, Y. Cpt: Efficient deep neural network training via cyclic precision. *arXiv preprint arXiv:2101.09868*, 2021.

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. A framework for few-shot language model evaluation, 07 2024. URL https://zenodo.org/records/12608602.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. arxiv. *arXiv preprint arXiv:2203.15556*, 2022.

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. *arXiv preprint arXiv:2001.08361*, 2020.

Kumar, T., Ankner, Z., Spector, B. F., Bordelon, B., Muennighoff, N., Paul, M., Pehlevan, C., Ré, C., and Raghunathan, A. Scaling laws for precision. *arXiv preprint arXiv:2411.04330*, 2024.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Liu, S.-y., Liu, Z., Huang, X., Dong, P., and Cheng, K.-T. Llm-fp4: 4-bit floating-point quantized transformers. *arXiv preprint arXiv:2310.16836*, 2023.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Micikevicius, P., Stosic, D., Burgess, N., Cornea, M., Dubey, P., Grisenthwaite, R., Ha, S., Heinecke, A., Judd, P., Kamalu, J., et al. Fp8 formats for deep learning. *arXiv preprint arXiv:2209.05433*, 2022a.

Micikevicius, P., Stosic, D., Burgess, N., Cornea, M., Dubey, P., Grisenthwaite, R., Ha, S., Heinecke, A., Judd, P., Kamalu, J., et al. Fp8 formats for deep learning. *arXiv preprint arXiv:2209.05433*, 2022b.

Nvidia. Nvidia blackwell architecture technical brief. https://resources.nvidia.com/en-us-blackwell-architecture.

NVIDIA. Transformer engine. https://github.com/NVIDIA/TransformerEngine.

Peng, H., Wu, K., Wei, Y., Zhao, G., Yang, Y., Liu, Z., Xiong, Y., Yang, Z., Ni, B., Hu, J., et al. Fp8-lm: Training fp8 large language models. *arXiv preprint arXiv:2310.18313*, 2023.

Radford, A. Improving language understanding by generative pre-training. 2018.

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., and Catanzaro, B. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021, 2020.

Sun, X., Choi, J., Chen, C.-Y., Wang, N., Venkataramani, S., Srinivasan, V. V., Cui, X., Zhang, W., and Gopalakrishnan, K. Hybrid 8-bit floating point (hfp8) training and inference for deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Wang, A. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Wang, N., Choi, J., Brand, D., Chen, C.-Y., and Gopalakrishnan, K. Training deep neural networks with 8-bit floating point numbers. *Advances in neural information processing systems*, 31, 2018.

Wang, R., Gong, Y., Liu, X., Zhao, G., Yang, Z., Guo, B., Zha, Z., and Cheng, P. Optimizing large language model training using fp4 quantization. *arXiv preprint arXiv:2501.17116*, 2025.

Weber, M., Fu, D., Anthony, Q., Oren, Y., Adams, S., Alexandrov, A., Lyu, X., Nguyen, H., Yao, X., Adams, V., et al. Redpajama: an open dataset for training large language models. *Advances in Neural Information Processing Systems*, 37:116462–116492, 2025.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Xi, H., Chen, Y., Zhao, K., TEH, K. J., Chen, J., and Zhu, J. Jetfire: Efficient and accurate transformer pretraining with int8 data flow and per-block quantization. In *Forty-first International Conference on Machine Learning*.

Xi, H., Li, C., Chen, J., and Zhu, J. Training transformers with 4-bit integers. *Advances in Neural Information Processing Systems*, 36:49146–49168, 2023.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. Opt: Open pre-trained transformer language models, 2022. *URL https://arxiv. org/abs/2205.01068*, 3:19–0, 2023.

A. FP4 Quantization

Quantization is the process of converting a data type with more bits (e.g., 32- or 16-bit floating points) into another data type with fewer bits (e.g., 4-bit floating points). In integer quantization, the real-valued variable X_R is quantized to an integer X_{INT} with the following formula:

$$X_{\text{INT}} = \alpha \left\lfloor \text{Clip}\left(\frac{X_R}{\alpha}, Q_{\text{min}}, Q_{\text{max}}\right)\right\rceil \tag{1}$$

Similar to integer quantization, in float point quantization, scaling and clipping of the values are required before quantization, as follows.

$$Q_{\text{max}} = -Q_{\text{min}} = (2 - 2^{-m})2^{2^e - b - 1} \tag{2}$$

$$\tilde{Q}_{\text{max}} = \alpha Q_{\text{max}} \tag{3}$$

$$X'_R = \text{Clip}\left(X_R, \tilde{Q}_{\text{min}}, \tilde{Q}_{\text{max}}\right) \tag{4}$$

Where the min/max value range of signed floating-point quantization can be calculated from Eq. (2), and the scaling factor α determines the quantization granularity. Thus, we can determine the upper and lower bounds of floating-point quantization and perform the clip operation accordingly.

After scaling and clipping, we can quantize the real value into a specific data format. Unlike INT quantization, floating-point numbers have different quantization levels for different values. Therefore, we first need to determine the quantization step size:

$$\tilde{\alpha} = 2^{-\tilde{b}} = 2^{-b} \cdot \alpha \tag{5}$$

$$v = \begin{cases} 2^{\lfloor \log_2 |X'_R|/\tilde{\alpha} \rceil - m} & \text{if } \lfloor \log_2 |X'_R|/\tilde{\alpha} \rceil \geq 1 \\ 2^{1-m} & \text{otherwise} \end{cases} \tag{6}$$

Here, the quantization level v is determined by $x'_R/\tilde{\alpha}$, after which the floating-point number can be quantized following the format of Eq. (8). A detailed explanation can be found in (Micikevicius et al., 2022a; Liu et al., 2023). Finally, the quantization formula can be expressed as follows:

$$X_{\text{FP}} = \tilde{\alpha} \cdot v \cdot \left\lfloor \frac{X'_R}{\tilde{\alpha} \cdot v} \right\rceil \tag{7}$$

B. Training Detail

We conducted our experiments on the Megatron (Shoeybi et al., 2019) framework and evaluated downstream performance using the transformers (Wolf et al., 2020) and lm-evaluation-harness (Gao et al., 2024), ensuring standardized and reproducible benchmarking. Hyperparameters remain consistent across precision settings for fair comparison. The learning rate follows a warm-up and cosine decay schedule, with the warm-up phase spanning 0.15% of total steps and the learning rate gradually decreasing to 10% of its peak over the remaining 90%. The peak learning rate is 1×10^{-4}, with a weight decay of 0.1 for Llama models. For GPT models, the peak learning rate is set to 6×10^{-4}, with a weight decay of 0.1 For the Adam optimizer, we use $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 1 \times 10^{-8}$. For llama models, the input sequences are fixed at 2048 tokens, and the batch size is 512, comprising approximately 1M tokens. For GPT models, the input sequences are fixed at 1024 tokens, and the batch size is 480, comprising approximately 0.5M tokens, detail model config can be found in Table 4.

Parameter	GPT-125M	GPT-335M	GPT-774M	LLaMA-125M	LLaMA-1B
Layers	12	24	36	12	48
Hidden Size	768	1024	1280	768	1280
Activation Function	GELU	GELU	GELU	SwiGLU	SwiGLU
Normalization	LayerNorm	LayerNorm	LayerNorm	RMSNorm	RMSNorm
FFN Hidden Size	3072	4096	5120	3072	3392
Sequence Length	1024	1024	1024	2048	2048
Attention Heads	12	16	20	12	20

Table 4. GPT and LLaMA Model Configurations

We quantize all the linear layers in the MLP and attention module to target precious, and leave multi-head attention and activation function in FP16 by employing FlashAttention (Dao et al., 2022). The master copy of the weights is kept in FP32. We quantize linear layers' inputs to target precious prior to each matmul, but leave layernorm's weight and bias to floating-point since they are relatively small.

For the calculation of theoretical computation cost, we first separately count the forward and backward computation amounts for each part of a Transformer block (considering only computations related to matrix multiplications, as these account for more than 95% of the total computation). Then, based on the assumptions that FP8 achieves twice the computation speed of FP16 and FP4 achieves four times the computation speed of FP16, we compute the theoretical time required for each matrix multiplication. Finally, we obtain the theoretical computation cost for each method.

Since the quantized weight $\tilde{w}$ is an estimate of w, We directly use a straight-through estimator (Bengio et al., 2013) directly passes the gradient of $\tilde{w}$ to w:

$$\nabla_{\mathbf{w}} \mathcal{L}(\tilde{\mathbf{w}}) \leftarrow \nabla_{\tilde{\mathbf{w}}} \mathcal{L}(\tilde{\mathbf{w}}).$$

In our experiments, we found that different models have varying precision requirements. For the GPT-125M model, applying a per-token and per-channel FP4 quantization strategy for both forward computation and weight gradient computation is feasible. The final results are shown in Table 1. The use of per-token and per-channel quantization is designed to better align with matrix multiplication rules, allowing for efficient implementation on accelerators. However,

for the GPT-335M model, the per-token and per-channel FP4 quantization strategy becomes ineffective as the data volume increases. In this case, switching to per-block FP4 quantization for weight gradient computation enables training to proceed, with the final results also presented in Table 1. For the GPT-770M model, the quantization strategy used for GPT-335M becomes ineffective as training progresses. At this point, modifying the forward computation to use per-block FP4 quantization while increasing the precision of weight gradient computation to FP8 ensures stable training. Additionally, we validated the feasibility of the GPT-770M quantization strategy (as we discuss in Section 3) on the LLaMA-125M and LLaMA-1B models. It can be anticipated that as model size and data volume continue to grow, the precision requirements for model training will become increasingly stringent(Kumar et al., 2024). Ensuring that FP4 can support long-term, large-scale pretraining of large models remains a key direction for our future work.