Tutorial Proposal: Efficient Inference for Large Language Models – Algorithm, Model, and System

Xuefei Ning[1], **Guohao Dai**[2,4], **Haoli Bai**[3], **Lu Hou**[3], **Yu Wang**[1], **Qun Liu**[3]
[1]Tsinghua University [2]Shanghai Jiao Tong University [3]Noah's Ark Lab, Huawei [4]Infinigence-AI
foxdoraame@gmail.com, daiguohao@sjtu.edu.cn, baihaoli@huawei.com
houlu3@huawei.com, yu-wang@tsinghua.edu.cn, qun.liu@huawei.com

1 Description

Background. Large Language Models (LLMs) have attracted significant attention from both academia and industry in recent years. They are revolutionizing many applications, including chatbots, content creation, scientific discovery, and so on, while also marking a potential step towards defining and realizing "artificial general Intelligence".

However, the inference of LLMs incurs high computational costs, memory access overhead, and extensive memory usage (Wan et al., 2023; Miao et al., 2023; Zhou et al., 2024), leading to inefficiencies in terms of latency, throughput, power consumption, and storage. This poses challenges on the deployment of LLMs on both the edge and cloud. For example, the storage requirements make it hard to run a 70B model on personal laptops, while low throughput can negatively impact the profitability of search engines. In summary, developing efficient inference techniques is critical for deployment and development of LLMs.

Tutorial aim and design. Our tutorial focuses on the increasingly important topic of ***Efficient Inference for LLMs*** and aims to *provide a systematic understanding of key facts and methodologies from a designer's perspective*. We start by introducing the fundamental concepts and mechanisms of modern LLMs, along with the relevant software and hardware. Following this, we formally define the efficiency optimization problem. To equip the audience with a designer's mindset, we will explain how to diagnose efficiency bottlenecks for a given workload on specific hardware. In particular, we will demonstrate how to use the basic theoretical roofline model and the NVIDIA toolchain to identify these bottlenecks.

With all the tools at our disposal, we will begin with a conceptual analysis of the key factors contributing to inefficiency (Zhou et al., 2024), namely the autoregressive sampling scheme, model size, and the core attention operator. Next, we will introduce our full-stack taxonomy of efficient inference methods for LLMs, as shown in Fig. 1. We will walk through each category of methodology, using one to three representative methods as examples for each leaf subcategory, elaborating on the design logic behind each method and which inefficiency factors they primarily address. Finally, we will wrap up with a few demonstrations, a takeaway summary, and future research directions.

Taxonomy. As shown in Fig. 1, we classify efficient inference methods into algorithm-, model- and system-level ones.
(1) *Algorithm-level optimization* includes efficient decoding methods, input compression methods, as well as alternative generative paradigms beyond the autoregressive model.
(2) *Model-level optimization* designs efficient model structures or cuts down model-level redundancy statically or dynamically.
(3) *System-level optimization* optimizes the inference engine or the serving system without altering the model computation graph.

Algorithm-level optimization methods.
Algorithm-level optimization supplements or modifies the sampling method beyond plain autoregressive sampling scheme.
(1) To mitigate the long latency and low hardware utilization resulting from the token-by-token autoregressive decoding method, *efficient decoding methods* focus on parallel generation, verification, or refinement strategies. Jacobi decoding methods (Santilli et al., 2023) transform the generation of multiple consecutive tokens into a non-linear equation system and use fixed-point iterations to iteratively refine tokens in parallel. Speculative decoding methods (Stern et al., 2018; Chen et al., 2023; Leviathan et al., 2023; Cai et al., 2024; Li et al., 2024b; Christopher et al., 2024) employ lightweight methods to propose candidate consecu-

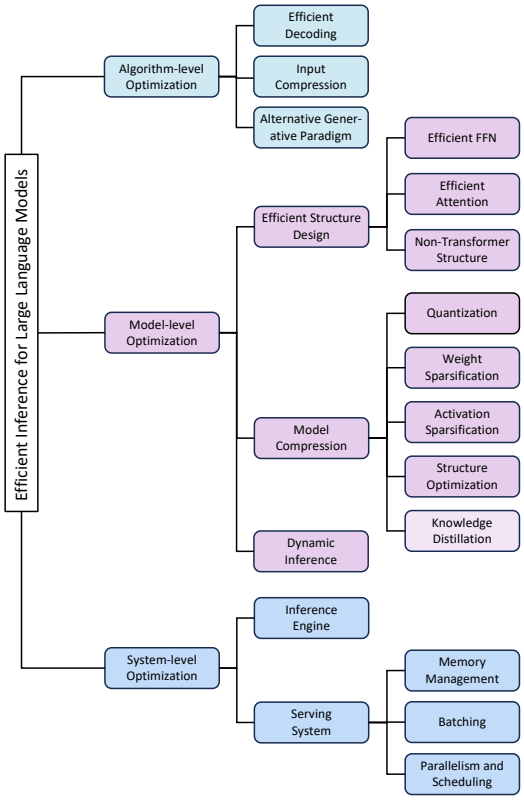

Figure 1: Taxonomy of efficient inference methods for LLMs. Modified for better clarity and organization from our previous survey (Zhou et al., 2024).

tive tokens and then use parallel verification with the original LLM to accept a subsequence. Agentic output organization methods (Ning et al., 2024; Liu et al., 2024d; Jin et al., 2024) leverage the planning ability of the LLM to plan the output structure and parallelly generate loosely related content parts.

(2) To process long inputs more efficiently, *input compression methods* focus on reducing the input context. Prompt compression (Chevalier et al., 2023; Li et al., 2023; Jiang et al., 2023a) and summary (Xu et al., 2024) methods compress the prompt either online or offline. Retrieval-augmented generation methods (Lewis et al., 2020; Gao et al., 2023; Jiang et al., 2023b) retrieve relevant information from external sources, avoiding including all information within the long context.

(3) Researchers are also exploring *alternative generative paradigms* (Wu et al., 2023; Lou et al., 2024) beyond standard autoregressive models. These models enable next-k-token or even randomized masked prediction, thereby addressing the inefficiency of token-by-token decoding.

Model-level optimization methods.

(1) *Efficient structure design* devises new neural network architectures, which often require retraining from scratch. These methods can be categorized into designing more efficient feed-forward networks (FFNs), such as the popular mixture-of-experts (MoE) design (Liu et al., 2024a,c; Jiang et al., 2024), more efficient attention mechanisms, including multi-query attention (Shazeer, 2019; Ainslie et al., 2023) and kernel-based attention (Katharopoulos et al., 2020; Peng et al., 2022), non-transformer alternatives like state-space models (SSMs) (Dao and Gu, 2024; Peng et al., 2023), and macro architecture design (Liu et al., 2024e; Tang et al., 2024b; Lieber et al., 2024).

(2) *Model compression* reduces the redundancy in the model's computation graph from multiple dimensions in a *static* manner, meaning that the computation graph for a given sequence of the same length is not content-dependent and is determined at deployment. These dimensions includes reducing the data precision (i.e., quantization) (Lin et al., 2024; Xiao et al., 2023; Liu et al., 2023; Dettmers et al., 2024), sparsifying attention, activation, and KV cache (Xiao et al., 2024; Fu et al., 2024), sparsifying weights (Frantar and Alistarh, 2023; Sun et al., 2023), and pruning the structure (Ma et al., 2023; Yuan et al., 2023; Liu et al., 2024b). Notably, some of these methods are designed to work with little or no retraining. We also briefly introduce *knowledge distillation* techniques (Gu et al., 2024), which transfer knowledge from a stronger model to a smaller one and can be combined with all the aforementioned compression approaches.

(3) *Dynamic inference* reduces the redundancy in the model's computation graph in a *dynamic* manner, meaning the computation graph for a given sequence of the same length is content-dependent and determined during inference. Many of the aforementioned dimensions can be optimized dynamically, such as sparsifying the KV cache (Zhang et al., 2024; Tang et al., 2024a), early exiting or layer skipping (Raposo et al., 2024; Schuster et al., 2022). Note that we categorize mixture-of-experts (MoE) under efficient structure design.

System-level optimization methods.

(1) *Inference engine techniques* encompass optimizations at both the graph level (Zhai et al., 2023; Hong et al., 2024) and operator level (Dao et al., 2022, 2023), as well as offloading strategies (Sheng et al., 2023).

(2) *Serving system* mainly aims to improve the handling of handling asynchronous requests, which

necessitates memory management optimizations to accomodate more requests (Kwon et al., 2023), efficient batching and scheduling strategies to improve the throughput (Yu et al., 2022), and specialized optimizations for distributed systems to better utilize distributed resources (Patel et al., 2024).

2 Tutorial Outline

2.1 Introduction (5 minutes)

- Background and motivation.
- Tutorial overview.

2.2 Preliminary and Problem Definition (15 minutes)

- Preliminary on mordern LLM.
- Preliminary on hardware architecture.
- Preliminary on mordern software-hardware stack at the cloud and edge.
- Application scenarios at the cloud and edge.
- The problem definition of efficiency optimization.

2.3 Bottleneck Diagnosis Tool From a Designer's Perspective (15 minutes)

- Introduce the concepts related to efficiency bottleneck analysis, including computation-bounded, memory-bounded and so on.
- Efficiency bottleneck diagnosis – the roofline method.
- Efficiency bottleneck diagnosis – taking the NVIDIA toolchain as an example.

2.4 Conceptual Analysis and Taxonomy of Efficient LLM Inference (5 minutes)

- Conceptual factors that cause the inefficiency of LLMs.
- Method taxonomy.

2.5 Model-level Optimization (40 minutes)

- Model compression: Quantization; Sparsification; Structure optimization, Knowledge distillation.
- Efficient structure design: MoE; SSM; Macro architecture design for smaller LLMs.
- Dynamic inference.

2.6 System-level Optimization (30 minutes)

- Efficient inferenece engine: Kernel design.
- Serving system: Continous batching; Scheduling strategy.

2.7 Algorithm-level Optimization (20 minutes)

- Efficient decoding algorithms: Speculative decoding; Agentic output organization.
- Input compression.
- Alternative generative paradigms.

2.8 Demonstrations (5 minutes)

2.9 Takeaways and Future Directions (10 minutes)

3 Tutorial Information

Type of the Tutorial. Cutting-edge.

Length. We plan to deliver a 3-hour tutorial, including breaks and time for Q&A.

Target Audience. Our tutorial is suitable for anyone with a basic understanding of machine learning, especially researchers or engineers who focused on improving the efficiency of LLMs. We aim to help our audiences build a comprehensive knowledge framework for efficient inference, introducing commonly used solutions and the latest advancements in this field.

Breadth. We estimate that less than 20% of the content presented in this tutorial will be the research from the tutor team, with the majority covering fundamental concepts and research from other scholars.

Diversity Considerations. The tutorial tutors come from four institutions, including two universities and two companies, and consist of young professors, senior professors, and industry researchers. Together, they offer insights on efficient inference from both academic and industry perspectives. Two female researchers are part of the tutor team. Their expertise covers diverse areas such as NLP algorithms, efficient algorithm design, software development, hardware design, and so on.

To encourage the participation of diverse audiences, we will advertise our tutorial through our website, social media, and push platforms in advance. In addition, we will reach out to our academic communities to encourage attendance.

Estimated Audience Size. Given that language models are now used in a range of NLP tasks and retrieval-based approaches have been applied to diverse domains, we estimate that the number of audiences will be around 150+.

Venues. We prefer ACL due to the growing interest in its efficiency-related tracks, as well as the travel and scheduling constraints of some of our tutors.

Technical Equipment. We need Internet access to show some demonstrations.

Open Access. We plan to make all teaching materials available online, including tutorial slides, tutorial code, demonstration videos, and tutorial videos, and allow them to be published in the ACL Anthology.

Ethical Considerations. A wide range of lossy model compression methods, such as quantization and pruning, can significantly enhance the inference efficiency and reduce carbon emissions. However, these methods may also lead to performance degradation. Previous research has demonstrated that lossy compression can, in some cases, cause "jailbreak" scenarios (Kumar et al., 2024; Li et al., 2024a), where large language models (LLMs) generate harmful responses. Therefore, it is crucial to consider the safety implications of efficient inference techniques.

Past Tutorials. A recent tutorial on similar topics is (Ren et al., 2023). However, it focuses more on efficient models for computer vision tasks, while this totorial pays more attention to large langauge models from the holistic view of algorithm, model, and system design.

In addition, the tutor team has published several surveys, tutorials, textbooks, and has organized workshops related to this topic.

- A survey on efficient LLM inference: A Survey on Efficient Inference for Large Language Models (Zhou et al., 2024).
- Some previous tutorial slides: Introduction on model compression; Introduction on LLM quantization.
- A series of NeurIPS workshops on Efficient Natural Language and Speech Processing.
- A textbook in Chinese, published by the Publishing House of Electronics Industry in 2024: "Efficient Deep Learning: Model Compression and Design".

4 Reading List

- A survey on efficient inference for large language models (Zhou et al., 2024).
- Eagle: Speculative sampling requires rethinking feature uncertainty (Li et al., 2024b).
- Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality (Dao and Gu, 2024).
- Smoothquant: Accurate and efficient post-training quantization for large language models (Xiao et al., 2023).
- Evaluating quantized large language models (Li et al., 2024a).
- Efficient streaming language models with attention sinks (Xiao et al., 2024).
- FlashAttention: Fast and memory-efficient exact attention with io-awareness (Dao et al., 2022).
- Efficient memory management for large language model serving with PagedAttention (Kwon et al., 2023).

Tutor Biography

Xuefei Ning is a research-track assistant professor with the Department of Electronic Engineering at Tsinghua University. She obtained her Ph.D. at Tsinghua University in 2021. Her research focuses on efficient deep learning. She has published 20+ papers on leading AI conferences and journals. She has published a Chinese book on efficient deep learning. She will serve as a senior area chair for ACL 2025, an area chair for CVPR 2025.

Guohao Dai is an associate professor with the Department of Electronic Information and Electrical Engineering at Shanghai Jiao Tong University. His research focuses on sparse computing, heterogeneous hardware computing, emerging hardware architecture, etc. He served as Co-Chair for the Ph.D. Forum at DAC 2024, TPC member for DAC 2024/DAC 2023/VLSID 2024. He received the Best Paper Award in ASP-DAC 2019, and Best Paper Nominations in DATE 2024/DATE 2023/DAC 2022/DATE 2018. He is the winner of the NeurIPS Billion-Scale Approximate Nearest Neighbor Search Challenge in 2021, and the recipient of the Outstanding PhD Dissertation Award of Tsinghua University in 2019.

Haoli Bai is a researcher at Huawei Noah's Ark Lab. He obtained his Ph.D. at the Chinese University of Hong Kong in 2021. His research focus is efficient deep learning with the purpose to minimize memory and computational requirements, particularly for large language models. He has published multiple research works on network quantization, pruning, and relevant topics, with applications on Huawei Ascend Chips and products. He obtained the ACML Best Student Paper Runner-up Award (2016), and has served as the PC member for top AI conferences (e.g., NeurIPS, ICML, ICLR).

Lu Hou is a researcher at Huawei Noah's Ark Lab. She obtained her Ph.D. from Hong Kong University of Science and Technology in 2019. Her research focuses on developing efficient deep learning models with lower memory and computation costs, especially for large pre-trained language and multimodal models. Her researches have been published at leading conferences (e.g., NeurIPS, ICML, ICLR, ACL, EMNLP) as well as been applied to various chips, products and LLMs at Huawei.

Yu Wang is a professor, an IEEE fellow, the chair of the Department of Electronic Engineering in Tsinghua University, the dean of the Institute for Electronics and Information Technology in Tianjin, and the vice dean of the School of Information Science and Technology in Tsinghua University. His research interests include the application specific heterogeneous computing, processing-in-memory, intelligent multi-agent system, and power/reliability aware system design methodology. He has published more than 90 journals (64 IEEE/ACM journals) and 270 conference papers in the areas of EDA, FPGA, VLSI Design, and Embedded Systems, with the Google Scholar citation over 22,000. He has received four best paper awards and 12 best paper nominations. He has been an active volunteer in the design automation, VLSI, and FPGA conferences. He is the co-founder of Deephi Tech (a leading deep learning solution provider), which is acquired by Xilinx (AMD) in 2018. He is also the promoter of Infinigence AI Tech (a leading AI infrastructure solution provider), which achieves industry-leading large language model inference performance on more than 10+ different chips.

Qun Liu is the chief scientist of Speech and Language Computing of Huawei Noah's Ark Lab. He is formerly a professor of Dublin City University, the Theme Learder of NLP at the ADAPT Centre, Ireland, a professor & researcher & the leader of NLP research group in the Institute of Computing Technology , Chinese Academy of Sciences (ICT-CAS). He obtained his B.Sc., M.Sc. and Ph.D. degrees in the University of Science and Technology of China, ICT-CAS, and Peking University respectively. His research interests cover natural language processing, language modeling, machine translation, question answering, dialog, etc. His academic achievements include ICTCLAS Chinese word segmentation and POS tagging system, syntax-based statistical machine translation, neural machine translation, machine translation evaluation, etc. He has been the leader or a participant in several large-scale projects funded by Chinese government, Irish government or European Union. He has published 300+ papers in academic conferences or journals, with 20,000+ citations. He has supervised 50+ Master or Ph.D. students into completion. He has obtained Google Research Award (2012), first prize of Qian Weichang Award for Chinese Information Processing Science and Technology (2010), and second prize of China National Award for Science and Technology Progress (2015), ACL Best Long Paper Awards (2019), and ACL Outstanding Paper Awards (2022, 2024).

References

Joshua Ainslie et al. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D Lee, Deming Chen, and Tri Dao. 2024. Medusa: Simple llm inference acceleration framework with multiple decoding heads. *arXiv preprint arXiv:2401.10774*.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. 2023. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*.

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. 2023. Adapting language models to compress contexts. *arXiv preprint arXiv:2305.14788*.

Jacob K Christopher, Brian R Bartoldson, Bhavya Kailkhura, and Ferdinando Fioretto. 2024. Speculative diffusion decoding: Accelerating language generation through diffusion. *arXiv preprint arXiv:2408.05636*.

Tri Dao and Albert Gu. 2024. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. In *Forty-first International Conference on Machine Learning*.

Tri Dao et al. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359.

Tri Dao et al. 2023. Flash-decoding for long-context inference. [Online]. https://crfm.stanford.edu/2023/10/12/flashdecoding.html.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Elias Frantar and Dan Alistarh. 2023. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning*, pages 10323–10337. PMLR.

Tianyu Fu, Haofeng Huang, Xuefei Ning, Genghan Zhang, Boju Chen, Tianqi Wu, Hongyi Wang, Zixiao Huang, Shiyao Li, Shengen Yan, et al. 2024. Moa: Mixture of sparse attention for automatic large language model compression. *arXiv preprint arXiv:2406.14909*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*.

Ke Hong, Guohao Dai, Jiaming Xu, Qiuli Mao, Xiuhong Li, Jun Liu, Yuhan Dong, Yu Wang, et al. 2024. Flashdecoding++: Faster large language model inference with asynchronization, flat gemm optimization, and heuristics. *Proceedings of Machine Learning and Systems*, 6:148–161.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023a. Llmlingua: Compressing prompts for accelerated inference of large language models. *arXiv preprint arXiv:2310.05736*.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*.

Shuowei Jin, Yongji Wu, Haizhong Zheng, Qingzhao Zhang, Matthew Lentz, Z Morley Mao, Atul Prakash, Feng Qian, and Danyang Zhuo. 2024. Adaptive skeleton graph decoding. *arXiv preprint arXiv:2402.12280*.

Angelos Katharopoulos et al. 2020. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR.

Divyanshu Kumar, Anurakt Kumar, Sahil Agarwal, and Prashanth Harshangi. 2024. Increased llm vulnerabilities from fine-tuning and quantization. *arXiv preprint arXiv:2404.04392*.

Woosuk Kwon et al. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pages 19274–19286. PMLR.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Shiyao Li, Xuefei Ning, Luning Wang, Tengxuan Liu, Xiangsheng Shi, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. 2024a. Evaluating quantized large language models. *arXiv preprint arXiv:2402.18158*.

Yucheng Li, Bo Dong, Chenghua Lin, and Frank Guerin. 2023. Compressing context to enhance inference efficiency of large language models. *arXiv preprint arXiv:2310.06201*.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. 2024b. Eagle: Speculative sampling requires rethinking feature uncertainty. *arXiv preprint arXiv:2401.15077*.

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, et al. 2024. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of Machine Learning and Systems*, 6:87–100.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. 2024a. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*.

Enshu Liu, Junyi Zhu, Zinan Lin, Xuefei Ning, Matthew B Blaschko, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. 2024b. Efficient expert pruning for sparse mixture-of-experts language models: Enhancing performance and reducing inference costs. *arXiv preprint arXiv:2407.00945*.

Liyuan Liu, Young Jin Kim, Shuohang Wang, Chen Liang, Yelong Shen, Hao Cheng, Xiaodong Liu, Masahiro Tanaka, Xiaoxia Wu, Wenxiang Hu, et al. 2024c. Grin: Gradient-informed moe. *arXiv preprint arXiv:2409.12136*.

Mingdao Liu, Aohan Zeng, Bowen Wang, Peng Zhang, Jie Tang, and Yuxiao Dong. 2024d. Apar: Llms can do auto-parallel auto-regressive decoding. *arXiv preprint arXiv:2401.06761*.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. 2023. Llm-qat: Data-free quantization aware training for large language models. *arXiv preprint arXiv:2305.17888*.

Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, et al. 2024e. Mobilellm: Optimizing sub-billion parameter language models for on-device use cases. *arXiv preprint arXiv:2402.14905*.

Aaron Lou, Chenlin Meng, and Stefano Ermon. 2024. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Forty-first International Conference on Machine Learning*.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36:21702–21720.

Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Hongyi Jin, Tianqi Chen, and Zhihao Jia. 2023. Towards efficient generative large language model serving: A survey from algorithms to systems. *arXiv preprint arXiv:2312.15234*.

Xuefei Ning, Zinan Lin, Zixuan Zhou, Zifu Wang, Huazhong Yang, and Yu Wang. 2024. Skeleton-of-thought: Prompting llms for efficient parallel generation. In *The Twelfth International Conference on Learning Representations*.

Pratyush Patel, Esha Choukse, Chaojie Zhang, Aashaka Shah, Íñigo Goiri, Saeed Maleki, and Ricardo Bianchini. 2024. Splitwise: Efficient generative llm inference using phase splitting. In *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*, pages 118–132. IEEE.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, et al. 2023. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*.

Hao Peng et al. 2022. Random feature attention. In *International Conference on Learning Representations*.

David Raposo, Sam Ritter, Blake Richards, Timothy Lillicrap, Peter Conway Humphreys, and Adam Santoro. 2024. Mixture-of-depths: Dynamically allocating compute in transformer-based language models. *arXiv preprint arXiv:2404.02258*.

Jian Ren, Sergey Tulyakov, and Ju Hu. 2023. Efficient neural networks: From algorithm design to practical mobile deployments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Andrea Santilli, Silvio Severino, Emilian Postolache, Valentino Maiorca, Michele Mancusi, Riccardo Marin, and Emanuele Rodolà. 2023. Accelerating transformer inference for translation via parallel decoding. *arXiv preprint arXiv:2305.10427*.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. 2022. Confident adaptive language modeling. *Advances in Neural Information Processing Systems*, 35:17456–17472.

Noam Shazeer. 2019. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*.

Ying Sheng, Lianmin Zheng, Binhang Yuan, Zhuohan Li, Max Ryabinin, Beidi Chen, Percy Liang, Christopher Ré, Ion Stoica, and Ce Zhang. 2023. Flexgen: High-throughput generative inference of

large language models with a single gpu. In *International Conference on Machine Learning*, pages 31094–31116. PMLR.

Mitchell Stern, Noam Shazeer, and Jakob Uszkoreit. 2018. Blockwise parallel decoding for deep autoregressive models. *Advances in Neural Information Processing Systems*, 31.

Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. 2023. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*.

Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. 2024a. Quest: Query-aware sparsity for efficient long-context llm inference. *arXiv preprint arXiv:2406.10774*.

Yehui Tang, Fangcheng Liu, Yunsheng Ni, Yuchuan Tian, Zheyuan Bai, Yi-Qi Hu, Sichao Liu, Shangling Jui, Kai Han, and Yunhe Wang. 2024b. Rethinking optimization and architecture for tiny language models. *arXiv preprint arXiv:2402.02791*.

Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, Mosharaf Chowdhury, et al. 2023. Efficient large language models: A survey. *arXiv preprint arXiv:2312.03863*, 1.

Tong Wu, Zhihao Fan, Xiao Liu, Hai-Tao Zheng, Yeyun Gong, Jian Jiao, Juntao Li, Jian Guo, Nan Duan, Weizhu Chen, et al. 2023. Ar-diffusion: Auto-regressive diffusion model for text generation. *Advances in Neural Information Processing Systems*, 36:39957–39974.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, pages 38087–38099. PMLR.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2024. Recomp: Improving retrieval-augmented lms with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*.

Gyeong-In Yu et al. 2022. Orca: A distributed serving system for transformer-based generative models. In *Proceedings of the 16th USENIX Symposium on Operating Systems Design and Implementation*, pages 521–538.

Zhihang Yuan, Yuzhang Shang, Yue Song, Qiang Wu, Yan Yan, and Guangyu Sun. 2023. Asvd: Activation-aware singular value decomposition for compressing large language models. *arXiv preprint arXiv:2312.05821*.

Yujia Zhai, Chengquan Jiang, Leyuan Wang, Xiaoying Jia, Shang Zhang, Zizhong Chen, Xin Liu, and Yibo Zhu. 2023. Bytetransformer: A high-performance transformer boosted for variable-length inputs. In *2023 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 344–355. IEEE.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. 2024. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36.

Zixuan Zhou, Xuefei Ning, Ke Hong, Tianyu Fu, Jiaming Xu, Shiyao Li, Yuming Lou, Luning Wang, Zhihang Yuan, Xiuhong Li, et al. 2024. A survey on efficient inference for large language models. *arXiv preprint arXiv:2404.14294*.