SCOPE: SCALABLE OPTIMIZATION FOR EFFICIENT AND ADPATIVE FOUNDATION MODELS

Souvik Kund[i], **Tianlong Chen**[n], **Shiwei Liu**[o], **Haizhong Zheng**[m],
Amir Yazdanbakhsh[d], **Beidi Chen**[c], **Yingyan (Celine) Lin**[g]
[i]Intel Labs USA, [n]UNC at Chapel Hill, [m]University of Michigan Ann Arbor, [o]University of Oxford UK,
[c]Carnegie Mellon University, [d]Google DeepMind USA, [g]Georgia Institute of Technology
souvikk.kundu@intel.com, tianlong@cs.unc.edu, shiwei.liu@maths.ox.ac.uk,
hzzheng@umich.edu, ayazdan@google.com , beidic@andrew.cmu.edu,
celine.lin@gatech.edu
Web: scope-iclr2025-workshop-submission

1 WORKSHOP SUMMARY

In the rapidly evolving landscape of AI, the development of scalable optimization methods to yield efficient and adaptive foundation models has significant demand in the space of their inference service. In specific, enabling model efficiency while allowing them to be adaptable to various new down-stream tasks has multifold challenges. *Firstly*, the model's ability to quickly learn adaptive and efficient sub-model selection on different tasks requires the capability to perform continual weight updates, compute- and memory-efficient fine-tuning, and personalized adaptation. *Secondly*, with the increased demand for long context understanding and reasoning, the model needs to yield such efficient adaptation with the informative usefulness of the query-specific token fetching. For instance, imagine a model that continually learns from current news events, adapting to the ever-changing global landscape by integrating up-to-date knowledge. Such models may not only need efficient fine-tuning to new incoming data stream, but also understand efficient handling of the KV cache that may keep on growing with the requirement to handle longer contextual information. Additionally, the integration of retrieval-augmented generation (RAG) into foundation models can ensure that generated content is not only relevant, but also reflects the most current knowledge while costing the prefill size to go up. *Thirdly,* with such growing demand for contextual adaptation, mixture of experts (MoE) models have also received significant traction that can perform test time adaptation via learned routing policy. In addition, the emergence of sub-quadratic models with constant KV states as opposed to KV caching of transformers, has opened up a new avenue of the model's adaptation ability in the context of information retention into compressive KV states. These capabilities rely on techniques for adapting foundation models, including fine-tuning, conversion, distillation, and in-context/few-shot learning. This workshop aims to capture **advances in scalable, adaptive fine-tuning, calibration, and conversion to yield inference efficient *quadratic and sub-quadratic* foundation models, focusing on methodologies across vision, language, and multi-modal domains**. Hosting this workshop at ICLR aligns with the conference's mission to advance the frontiers of machine learning. The workshop aims to bring together interdisciplinary researchers from core ML/DL, efficient ML, computer vision, and NLP.

Activities. The workshop will feature diverse activities, including keynotes, a panel discussion, oral and poster sessions. We welcome high-quality original papers in the following two tracks:

- **Short/tiny paper track:** with a maximum limit of 2 pages (as per the guideline of 2024 ICLR tiny paper track), without references and appendix.
- **Main paper track:** with a maximum limit of 5 pages, without references and appendix.

Outstanding papers will be selected for oral presentation, while all accepted papers will be allowed to be presented as posters. The workshop organizers, speakers, and reviewers will abide by the ICLR rules regarding conflicts of interest. The sessions will be live-streamed via Zoom and recorded for offline viewing. Standard conference presentation equipment will be required.

Attendance Estimate. Based on prior ICLR and NeurIPS workshops on similar themes and topics, such as adaptive foundation models (AFM), R0-FoMo, we currently estimate between 70-100 paper submissions, and around 175-200 attendees.

Plan to Get Audience for Workshop. We plan to broadcast the workshop details at social media platforms, including X and LinkedIn, upon acceptance. Additionally, we plan to send short online flyers in email to our respective institutions (Industry and Academic) and encourage our technical program committee members to do the same for broader outreach and participation.

We have created a website for the workshop: scope-workshop.github.io.

2 Topics

- **Efficient Long Context Understanding.** Adaptation to efficient enablement of long context understanding ability with transformer based (non-compressive memory) (Tang et al., 2024; Chen et al., 2024; Jin et al., 2024; Jiang et al., 2024) and sub-quadratic (compressive memory) (Ben-Kish et al., 2024) foundation models.
- **Sub-Quadratic Models for Foundational Tasks and Personalization.** Understanding the efficiency and limitations of sub-quadratic models as a potential alternative to transformer based LLMs/VLMs.
- **Quadratic to Sub-Quadratic Model Conversion.** Understanding the difference in attention understanding patterns between quadratic and sub-quadratic (Mamba2 (Dao & Gu, 2024), Griffin (De et al., 2024)) models and learn efficient conversion and distillation (Wang et al., 2024b) strategy from the earlier to the later to have inference time throughput advantage.
- **Task Specific Adaptive Foundation Models.** Different techniques for customizing models to individual user preferences, tasks, or domains, ensuring more relevant and effective interactions (Salemi et al., 2023; Zhang et al., 2023).
- **Retrieval Augmented Generation for Efficient Contextual Processing.** Integration of external knowledge sources to enhance the long context generation capabilities of models (Li et al., 2024).
- **Efficient Sub-Quadratic Foundation Models.** Post training optimization methods to yield compute and memory efficient sub-quadratic model (Pierro & Abreu, 2024; Shukla et al., 2024).
- **Adaptive Fine-Tuning for Multimodal Foundation Models.** Techniques for leveraging data from multiple modalities (e.g., text, images, robot interactions) into a unified framework (Yang et al., 2024; Xu et al., 2023).
- **Efficient Fine-Tuning for Continual Adaptation and Personalization.** Techniques and challenges in updating model weights continually to adapt to new information without forgetting previously learned knowledge. Further scope includes OOD generalization via memory efficient fine-tuning (Mao et al., 2024; Chen et al., 2023; Azizi et al., 2024) and calibration.
- **Model Optimization for Latency and Throughput Efficient Inference.** Improving model serving efficiency via post training optimization of foundation models (You et al., 2024; Kang et al., 2024; Cai et al., 2024) for various language and vision applications.
- **Adaptive Routing with Mixture of Experts.** Adaptive routing and sparse expert selection for mixture of experts (Sukhbaatar et al., 2024) and mixture of agents (Wang et al., 2024a).

3 Invited Speakers

† denotes a confirmed speaker, ○ denotes a speaker who has been invited.

Kate Saenko (She/her).$^\circ$ is a Professor of Computer Science at Boston University and the Director of the Computer Vision and Learning Group. Her focus is on adaptive vision and ML. She has authored Cola: A Benchmark for Compositional Text-to-Image Retrieval (NeurIPS 2023), among other papers exploring the personalization potential of diffusion models. Scholar: Kate-scholar, Homepage: Kate-home.

Soham De (He/him).$^\circ$ is a Research Scientist at DeepMind in London, where he works on better understanding and improving large-scale deep learning. He currently focuses on topics in optimization and initialization. He has been the lead or part of Google's recent pioneering projects on new efficient foundation models development including Griffin, Gemma, and RecurrentGemma. Scholar: Soham-scholar, Homepage: Soham-home.

Yu Cheng (He/him).[†] is a Professor in Computer Science and Engineering at the Chinese University of Hong Kong. From 2018-2023, he was a Principal Researcher at Microsoft Research Redmond. His research interests lie in model compression and efficiency, deep generative models, and large multimodal/language models. From 2021 to 2023, he led several teams to productize these techniques for Microsoft-OpenAI core models (Copilot, DALL-E-2, ChatGPT, GPT-4). Many of his paper have won the the best paper awards including Cybersecurity Best Paper 2024 and the Outstanding Paper Award at NeurIPS 2023. Scholar: Yucheng-scholar, Homepage: Yucheng-home.

Christopher Ré (He/him).[°] is an associate professor in the Stanford AI Lab (SAIL) at Stanford University. Ré 's works got multiple best paper and test of time awards at to-tier conferences like ICML, NeurIPS, and MIDL. His recent works on sub-quadratic state space models (SSMs) as a potential replacement of the quadratic transformer blocks has played a key role in the development of new Mamba foundation models. Ré has also won the Alfred P. Sloan Research Fellowship and Robert N. Noyce Faculty Fellowship both in 2013. Scholar: Chris-scholar, Homepage: Chris-home.

Zechun Liu (She/her)[†] is a senior research scientist at Meta Reality Labs. At Meta she has been leading various efforts in efficient pre-training, fine-tuning, and inference of foundation models. Her notable works including the recent effort for on-device LLM serving (MobileLLM, LLM-QAT, and SpinQuant) have a cumulative citation count of more than 4,600. She has also co-organized successful workshops on efficient training/inference at ICCV'23 and CVPR'21. Scholar: Zechun-scholar, Homepage: Zechun-home.

Zhangyang (Atlas) Wang (He/him)[†] is a tenured Associate Professor at The University of Texas at Austin. He is currently the full-time Research Director for XTX Markets, heading their new AI Lab in New York City. His recent core research mission is to leverage, understand, and expand the role of low dimensionality in ML and optimization, whose impacts span over many important topics such as the efficiency and trust issues in large language models (LLMs) as well as generative vision. He co-founded the new Conference on Parsimony and Learning (CPAL) and is its inaugural Program Chair. He is an elected technical committee member of IEEE MLSP and IEEE CI; and regularly serves as (senior) area chair, invited speakers, tutorial/workshop organizers, various panelist positions, and reviewers. He is an ACM Distinguished Speaker, an IEEE senior member, and has been AI's top 10 to watch (2023). Scholar: Atlas-scholar, Homepage: Atlas-home.

Zhuang Liu (He/him)[†] is a Research Scientist at Meta Fundamental AI Research (FAIR), New York. HIs primary research areas are deep learning and computer vision. He works on deep learning model architectures, training, efficiency, and understanding. He has been lead or co-lead of various impactful works like Densely connected CONVNet ($>$ 48900 citations, CVPR 2017), ConvNet for 2020s (CVPR 2022), and simple pruning for LLMs (ICLR 2024). Scholar: Zhunag-scholar, Homepage: Zhuang-home.

4 Tentative Schedule and Conflict Management

Timeline for Submissions. The workshop paper submission deadline will be **February 3, 2025**, AoE; accept/reject notifications will be sent by **March 5, 2025**, AoE; and the workshop will take place in person on April 27 or 28, 2025 (full-day). Table 1 outlines the tentative schedule of the workshop program. Import date to ICLR: 27 March 2025, 11.59 pm AoE.

Conflict Management. We will put utmost care to manage conflicts of interest in assessing submitted contributions. We have kept a diverse list of organizer and initial technical program committee list to ensure seamless dealing of this due to sufficient institutional diversity.

5 Diversity Commitment

At the heart of our workshop lies a strong commitment to diversity and inclusion. To encourage the attendance of diverse participants with various perspectives, we have broadened workshop topics by

Table 1: Proposed Schedule.

Time	Event
8:30-9:00	***Welcome and Keynote:*** Christopher Ré
9:00-9:30	***Invited talk:*** Kate Saenko
9:30-10:00	***Invited talk:*** Zhuang Liu
10:00-10:15	**Coffee Break**
10:15-10:45	***Contributed talks:*** Workshop Oral Papers I
10:45-11:45	Morning Poster Session
11:45-13:00	**Lunch Break**
13:00-13:30	***Keynote:*** Zhangyang (Atlas) Wang
13:30-14:00	***Invited talk:*** Yu Cheng
14:00-15:00	***Contributed talks:*** Workshop Oral Papers II
15:00-15:15	**Coffee Break**
15:15-15:45	***Invited talk:*** Soham De
15:45-16:15	***Invited talk:*** Zechun Liu
16:15-16:45	***Panel Discussion***
16:45-17:30	Closing Remarks and Afternoon Poster Session

stressing the importance of efficient fine-tuning and adaption for downstream tasks, efficient inference with improved contextual understanding, and beyond quadratic transformer-based architectures for foundation models.

Our organizing committee and speaker lineup were intentionally curated to reflect substantial diversity. The organizers originate from esteemed **affiliations** in both the industry (Google DeepMind, Intel Labs, Meta) and academia (Oxford, CMU, Georgia Tech, UNC-Chapel Hill). They represent a breadth of **geographies** from the United States, Europe, and Asia, showcasing a spectrum of **professional ranks** from (Staff) Research Scientists to faculty members, including Research Fellows, Assistant Professors, and Full Professors. Furthermore, our team of organizers and speakers is **gender-diverse**, featuring four female members and encompasses a variety of cultural backgrounds spanning the US, Asia, and Europe.

To support the growth of emerging scholars, our workshop will offer a Best Student Paper Award in addition to the standard Best Paper Award, aimed at recognizing and encouraging innovation among young researchers. We are also actively pursuing sponsorship opportunities to provide travel grants for students and individuals from underrepresented and marginalized communities, further promoting inclusivity and participation in cutting-edge AI research.

6 Related Workshops

• **NeurIPS 2024 Workshop on Adaptive Foundation Models (AFM):** Aiming to explore advances in adaptive foundation models, focusingon methodologies across vision, language, and multi-modal domains.

• **ICLR 2024 Practical ML for Low Resource Settings:** Aiming to foster collaborations and build a cross-domain community by featuring invited talks, panel discussions, contributed presentations (oral and poster) and round-table mixers.

• **NeurIPS 2023 Workshop on Robustness of Zero/Few-shot Learning in Foundation Models (R0- FoMo):** Focused on the robustness of few-shot and zero-shot learning, R0-FoMo examined the theoretical and empirical aspects of applying these methodologies in large foundation models. Our proposed workshop extends these topics by focusing on how adaptive foundation models can leverage few-shot and in-context learning for more efficient and personalized adaptation.

• **ICML 2023 Efficient Systems for Foundation Models (ES-FoMo):** addressed the challenges of scaling foundation models in compute, memory, and energy efficiency. Our workshop intends to build on these discussions by emphasizing the adaptive aspects, focusing on continual updates and personalization.

• **NeurIPS 2024 Edge-LLM Challenge**: Aiming to present novel deployment solutions for LLM at the edge. This is the first workshop on LLM's on-device deployment at NeurIPS 2024.

7 Organizers

Souvik Kundu (He/him).(Designated contact) Home: Souvik-home, Google Scholar: Souvik-scholar

Souvik Kundu is a Staff Research Scientist at Intel Labs, USA. Earlier he received his Ph.D. from the University of Southern California, USA. At Intel he has been the co-founder and co-lead of the research team focusing on scalable and novel AI primitives with research focusing on efficient inference and fine-tuning of foundation models. Specifically, his team explores designs, and optimizes algorithms for foundation models that can optimally run on existing hardware to meet reduced memory, compute, and latency demand. He has over 65 top-venue publications with multiple oral, Spotlight, and best-paper recommendations at various top-tier conferences. He was selected as one of the youngest outstanding liaison award winners (2023) from the Semiconductor Research Corporation (SRC), USA. He has previous organizing experience on a special session at ICASSP'24, along with the inaugural program committee experience at the Conference on Parsimony and Learning (CPAL)'24. He will be serving as an Industry Liaison Chair at CPAL'25.

Tianlong Chen (He/him).(Designated contact) Home: Tianlong-home, Google Scholar: Tianlong-scholar

Tianlong Chen is currently an Assistant Professor in the Department of Computer Science, at University of North Carolina at Chapel Hill. Earlier he received his Ph.D. degree in Electrical and Computer Engineering from University of Texas at Austin, TX, USA, in 2023. His research focuses on building accurate, trustworthy, and efficient machine learning systems. He is the recipient of the Cisco Faculty Award, OpenAI Researcher Access Award, Gemma Academic Program GCP Credit Award, IBM Ph.D. Fellowship, Adobe Ph.D. Fellowship, Graduate Dean's Prestigious Fellowship, AdvML Rising Star Award, and the Best Paper Award from the inaugural Learning on Graphs (LoG) Conference 2022. He has co-organized several tutorials in ICASSP'24, AAAI'24, and ICML'24. He has previous organizing experience as he was the inaugural program committee member at the Conference on Parsimony and Learning (CPAL)'24.

Shiwei Liu (He/him).(Designated contact) Home: Shiwei-home, Google Scholar: Shiwei-scholar

Shiwei Liu is a Royal Society Newton International Fellow at the University of Oxford. He was a Postdoctoral Fellow at the University of Texas at Austin. He obtained his Ph.D. with the Cum Laude from the Eindhoven University of Technology in 2022. His research goal is to leverage, understand, and expand the role of sparsity/low-rank in neural networks, whose impacts span many important topics, such as efficient training/inference/transfer of large-foundation models, robustness and trustworthiness, and generative AI. His current research interest focuses on improving the efficiency and accessibility of LLMs, making them accessible tooling to everyone. Dr. Liu has received rising star awards from KAUST and the Conference on Parsimony and Learning (CPAL). His Ph.D. thesis received the 2023 Best Dissertation Award from Informatics Europe. He has been the lead organizer for the "Edge LLM Challenge" at NeurIPS'24, with additional organization experience on "Neural Network Sparsity" workshop at ICLR'23.

Haizhong Zheng (He/him).(Designated contact) Home: Haizhong-home, Google Scholar: Haizhong-scholar

Haizhong Zheng is joining Carnegie Mellon University as a Postdoctoral researcher. He obtained his Ph.D. from the University of Michigan. His research interests are at the intersection of systems and machine learning, and his research goal is to bridge the gap between the rapid scaling of models and the slower scaling in hardware and high-quality data. More specifically, he has focused on developing algorithms to train models to utilize hardware resources more efficiently and designing algorithms to improve data efficiency for deep learning. His research has appeared at top conferences like ICLR, NeurIPS, CVPR, ECCV, and NDSS.

Amir Yazdanbakhsh (He/him). Home: Amir-home, Google Scholar: Amir-scholar

Amir Yazdanbakhsh is a Research Scientist at Google DeepMind, USA. He is the co-founder and co-lead of the Machine Learning for Computer Architecture team. His team works on research and development of machine learning algorithms and architectures to shape the future generation of Google ML accelerators. Earlier Amir received his Ph.D. degree in computer science from the Georgia Institute of Technology with work recognized by various awards, including Microsoft PhD Fellowship and Qualcomm Innovation Fellowship. For his pioneering research contributions in the field of efficient and scalable architecture and system design, in 2023 Amir has been inducted to the ISCA hall of fame. He has co-organized various workshops at various top-tier venues including the recently concluded "MLArchSys Workshop" hosted at ISCA 2024.

Beidi Chen (She/her). Home: Beidi-home, Google Scholar: Beidi-scholar

Beidi Chen is an Assistant Professor at Carnegie Mellon University and a Research Scientist at FAIR. Before that, she was a postdoctoral scholar at Stanford University. She received her Ph.D. from Rice University. Her research focuses on efficient deep learning; specifically, she designs and optimizes randomized algorithms on current hardware to accelerate large machine learning systems. Her work has won best paper runner-up at ICML 2022 and she was selected as a Rising Star in EECS by MIT and UIUC. She was a workshop chair for MLSys 2023 and 2024, and she co-organized many workshops at ICML and NeurIPS.

Yingyan (Celine) Lin (She/her). Home: Celine-home, Google Scholar: Celine-scholar

Yingyan (Celine) Lin is currently an Associate Professor in the School of Computer Science at the Georgia Institute of Technology. She leads the Efficient and Intelligent Computing (EIC) Lab, focusing on developing efficient machine learning solutions through cross-layer innovations, from AI algorithms to hardware accelerators and chip design. Her work aims to promote green AI and ubiquitous AI-powered intelligence. She received her Ph.D. in Electrical and Computer Engineering from the University of Illinois at Urbana-Champaign in 2017. Celine has received several prestigious awards, including the NSF CAREER Award (2021) and ACM SIGDA Outstanding Young Faculty Award (2022). In 2024, she received the SRC Young Faculty Award. She has co-organized various workshops at various top-tier venues including the recently concluded "MLArchSys Workshop" hosted at ISCA 2024.

8 Technical Program Committee

Reviewers (To be invited). Haoran You (Georgia tech), Wenhao Zheng (UNCCH), Sihwan Park (KAIST), June Yong Yang (KAIST), Sharath Nittur Sridhar (Intel), Zhenyu Zhang (UT Austin), Akshat Ramachandran (Georgia Tech), Hao Kang (Georgia Tech), Shamik Kundu (Intel), Pingzhi Li (UNCCH), Xinyu Zhao (UNCCH), Sukwon Yun (UNCCH), Huaizhi Qu (UNCCH), Mufan Qiu (UNCCH), Rana Muhammad Shahroz Khan (UNCCH), Ruichen Zhang (UNCCH), Lu Yin (U Surrey), Ajay Jaiswal (UT Austin), Qiao Xiao (TU Eindhoven), Boquan Wu (University of Luxembourg), Zhuoming Chen (CMU), Xinyu Yang (CMU), Yang Zhou (CMU), Ranajoy Sadhukhan (CMU), Hanshi Sun (CMU), Zeyu Liu (USC), SeyedArmin Azizi (USC), Sanjay Das (UTD), Sean Mcpherson (Intel), Ruisi Cai (UT Austin), Zhifan Ye (Georgia tech), Abhimanyu Bambhaniya (Georgia tech), Zheyu Shen (UMD), Jian Meng (Cornell), In Gim (Yale).

References

Seyedarmin Azizi, Souvik Kundu, and Massoud Pedram. Lamda: Large model fine-tuning via spectrally decomposed low-dimensional adaptation. *arXiv preprint arXiv:2406.12832*, 2024.

Assaf Ben-Kish, Itamar Zimerman, Shady Abu-Hussein, Nadav Cohen, Amir Globerson, Lior Wolf, and Raja Giryes. Decimamba: Exploring the length extrapolation potential of mamba. *arXiv preprint arXiv:2406.14528*, 2024.

Ruisi Cai, Saurav Muralidharan, Greg Heinrich, Hongxu Yin, Zhangyang Wang, Jan Kautz, and Pavlo Molchanov. Flextron: Many-in-one flexible large language model. *arXiv preprint arXiv:2406.10260*, 2024.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*, 2023.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. Longlora: Efficient fine-tuning of long-context large language models. In *The International Conference on Learning Representations (ICLR)*, 2024.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

Soham De, Samuel L Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, et al. Griffin: Mixing gated linear recurrences with local attention for efficient language models. *arXiv preprint arXiv:2402.19427*, 2024.

Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H Abdi, Dongsheng Li, Chin-Yew Lin, et al. Minference 1.0: Accelerating pre-filling for long-context llms via dynamic sparse attention. *arXiv preprint arXiv:2407.02490*, 2024.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. Llm maybe longlm: Self-extend llm context window without tuning. *ICML*, 2024.

Hao Kang, Qingru Zhang, Souvik Kundu, Geonhwa Jeong, Zaoxing Liu, Tushar Krishna, and Tuo Zhao. Gear: An efficient kv cache compression recipefor near-lossless generative inference of llm. *arXiv preprint arXiv:2403.05527*, 2024.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. Retrieval augmented generation or long-context llms? a comprehensive study and hybrid approach. *arXiv preprint arXiv:2407.16833*, 2024.

Yuren Mao, Yuhang Ge, Yijiang Fan, Wenyi Xu, Yu Mi, Zhonghao Hu, and Yunjun Gao. A survey on lora of large language models. *arXiv preprint arXiv:2407.11046*, 2024.

Alessandro Pierro and Steven Abreu. Mamba-ptq: Outlier channels in recurrent large language models. *arXiv preprint arXiv:2407.12397*, 2024.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. Lamp: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*, 2023.

Abhinav Shukla, Sai Vemprala, Aditya Kusupati, and Ashish Kapoor. Matmamba: A matryoshka state space model. *arXiv preprint arXiv:2410.06718*, 2024.

Sainbayar Sukhbaatar, Olga Golovneva, Vasu Sharma, Hu Xu, Xi Victoria Lin, Baptiste Rozière, Jacob Kahn, Daniel Li, Wen-tau Yih, Jason Weston, et al. Branch-train-mix: Mixing expert llms into a mixture-of-experts llm. *arXiv preprint arXiv:2403.07816*, 2024.

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

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024a.

Junxiong Wang, Daniele Paliotta, Avner May, Alexander M Rush, and Tri Dao. The mamba in the llama: Distilling and accelerating hybrid models. *arXiv preprint arXiv:2408.15237*, 2024b.

Jinjin Xu, Liwu Xu, Yuzhe Yang, Xiang Li, Yanchun Xie, Yi-Jie Huang, and Yaqian Li. u-llava: Unifying multi-modal tasks via large language model. *arXiv preprint arXiv:2311.05348*, 2023.

Yijun Yang, Tianyi Zhou, Kanxue Li, Dapeng Tao, Lusong Li, Li Shen, Xiaodong He, Jing Jiang, and Yuhui Shi. Embodied multi-modal agent trained by an llm from a parallel textworld. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26275–26285, 2024.

Haoran You, Yipin Guo, Yichao Fu, Wei Zhou, Huihong Shi, Xiaofan Zhang, Souvik Kundu, Amir Yazdanbakhsh, and Yingyan Lin. Shiftaddllm: Accelerating pretrained llms via post-training multiplication-less reparameterization. *arXiv preprint arXiv:2406.05981*, 2024.

Kai Zhang, Fubang Zhao, Yangyang Kang, and Xiaozhong Liu. Memory-augmented llm personalization with short-and long-term memory coordination. *arXiv preprint arXiv:2309.11696*, 2023.