Workshop on Sparsity in LLMs (SLLM)

Deep Dive into Mixture of Experts, Quantization, Hardware, and Inference

Organizers: Tianlong Chen [1] Utku Evci [2] Yani Ioannou* [3] Berivan Isik [2] Shiwei Liu [4] Adnan Mohammed [3,7] Aleksandra Nowak [5] Ashwinee Panda* [6]

Affiliations: [1]University of North Carolina at Chapel Hill [2]Google [3]University of Calgary [4]University of Oxford [5]Jagiellonian University [6]University of Maryland [7]Vector Institute

I. Workshop Summary

Large Language Models (LLMs) have emerged as transformative tools in both research and industry, excelling across a wide array of tasks. However, their growing computational demands—especially during inference—raise significant concerns about accessibility, environmental sustainability, and deployment feasibility. At the same time, sparsity-based techniques are proving critical not just for improving efficiency but also for enhancing interpretability, modularity, and adaptability in AI systems.

This workshop aims to bring together researchers and practitioners from academia and industry who are advancing the frontiers of sparsity in deep learning. Our scope spans several inter-related topics, including Mixture of Experts (MoEs), LLM inference and serving, network pruning, sparse training, distillation, activation sparsity, low-rank adapters, hardware innovations and quantization. A key objective is to foster connections and unlock synergies between traditionally independent yet highly related research areas, such as activation sparsity and sparse autoencoders (SAEs), or quantization and KV cache compression. Rather than focusing solely on efficiency, we aim to explore how sparsity can serve as a unifying framework across multiple dimensions of AI—driving advances in interpretability, generalization, and system design.

By facilitating the fusion of ideas from different topics, the workshop will create new opportunities for innovation. We encourage participants to think beyond traditional constraints, exploring how different forms of sparsity can inform each other and yield new algorithms. Whether the goal is faster inference, modular architectures, or more interpretable models, our aim is to catalyze research that deepens the integration of sparsity within AI.

Topics of interest include, but are not limited to:

- **(1) Mixture of Experts (MoEs) and Modularity.** Large-scale pretraining depends on increasing the number of parameters in a model, which improves sample efficiency and reduces the amount of training needed to achieve the same performance as smaller models. [1], [2]. The dominant architecture in this field is a dense Transformer [3], known for its efficient scaling with both parameter size and data. However, many industry deployments [4]–[9] have adopted a sparsely activated Mixture-of-Experts (MoE) Transformer architecture [10], as MoEs have been shown to scale even more effectively than dense Transformers [11]–[14]. MoEs learn a *routing* function that selectively activates the top-K subset of their modules (called *experts*), most relevant to a given input. This conditionally sparse activation [15], [16] allows to multiplicatively increase the model parameter count without significantly increasing the cost of training or inference.
- **(2) Parameter Sparsity/Pruning.** Existing efficient deep learning approaches based on sparsity, such as dynamic sparse training and pruning, have found great success in accelerating the inference of deep neural networks, in particular large Convolutional Neural Networks (CNNs) for computer vision tasks. However, despite this success, these approaches do not see the same success in accelerating LLMs, without significant adaption, as early work has shown. Seminal works, such as SparseGPT [17] and Wanda [18], demonstrated that LLMs can be pruned after training. However, the performance of the pruned/compressed model drastically drops beyond 50% sparsity. Moreover, recent work has shown that most of the pruning methods perform poorly when evaluated on more complex datasets/benchmarks [19]. Thus, post-training of LLMs is still an active area of research and further development is required to understand how to prune/compress LLMs without impacting their generalization.
- **(3) Interaction w/ Quantization and Distillation.** Emerging techniques that enhance efficiency often complement existing methods rather than replace them. For instance, quantization of parameters or activations can achieve compression ratios of up to 16x [20], [21], while model distillation is frequently employed to leverage pre-trained models in new training cycles [22]. Given the difficulty of supplanting these established methods, it is crucial to ensure that new techniques integrate seamlessly with them. Additionally,

*Points of Contact: Yani Ioannou, Ashwinee Panda

exploring existing methods and combining them with innovative approaches like sparsity can open up exciting new avenues for research and application.

(4) Activation Sparsity for Inference. Activation sparsity is a powerful tool for improving the efficiency of inference. Sparsity can be achieved during inference by modifying the activation function [23], [24] or predicting what neurons should be sparsified [25]. If LLMs can be sparsely activated [26] then the input/output (IO) operations can be reduced, and more zero elements in matmuls naturally reduces compute and power usage.

(5) Sparsity in Attention. The attention mechanism has become the bottleneck in training and serving LLMs, especially with long-context models designed to operate over very large Key-Value (KV) Caches. A number of methods have been proposed that seek to shift attention from an exact global mechanism to something more approximate. These methods include sparse attention [27], KV cache pruning [28], [29], KV cache quantization [30]–[32] that seek to cap the size of the KV cache, and various hybrid architectures [33]. In input space, papers have proposed compressing the prompt itself into the most relevant information [34]–[36] which will naturally decrease the size of the KV cache.

(6) Sparsity for Interpretability. One of the foremost directions in interpreting LLMs via mechanistic interpretability is the development of sparse autoencoders (SAEs). SAEs can be used to do post-hoc interpretation of directions in the activation space of an LLM [37], [38] by decomposing those activations in terms of a large dictionary of building block features. In order for a decomposition to be interpretable, it must be *sparse*—we should obtain low reconstruction error with just a few elements of the dictionary. Many papers have made an effort to scale up SAEs [39], [40] and recently Rajamanoharan, Lieberum, Sonnerat, *et al.* [41] showed that by applying a naturally sparse activation function, they can obtain sparse decompositions. We believe that providing a venue for the intersection of mechanistic interpretability and sparsity can spark more ideas to combine well-known ideas in the sparsity literature with SAEs.

(7) Hardware Innovation for Sparsity. Exploiting sparsity at the hardware level is not an easy feat, and a great deal of the growing success in sparsity in all areas of LLM development is due to better hardware support for sparsity on the latest GPUs [42] that makes its way into the most commonly used research software [43]. Academic-industry collaborations have found ways to efficiently find sparsity patterns [44], and move larger tensors into memory [45]. Innovative chip designs, such as those developed by Cerebras [46], that support unstructured sparsity, are pivotal in fully leveraging the potential of sparsity in the era of LLMs.

(8) Parameter Efficient Fine Tuning. Parameter Efficient Fine Tuning (PEFT) aims to optimize a pre-trained model for a downstream task while minimizing the number of parameters that need to be trained, at the same time maintaining (or surpassing) the performance level of a fully fine-tuned model [47]. PEFT approaches provide exciting opportunities to reuse a large pre-trained model by focusing on ideas of flexibility and modularity. Common approaches in this field include inserting small adjustable modules, often utilizing low-rank projections such as Adapters or LoRA methods [48], [49], as well as adding trainable prefixes to the inputs of intermediate layers in the network [50], [51]. Other alternatives focus on optimizing a specific set of weights [52] or incorporate a separate side network alongside the pre-trained model to facilitate efficient gradient propagation, thereby reducing both memory consumption and training time [53], [54].

II. Workshop Format / Tentative schedule

A. Submission Tracks

We will host two submission tracks on OpenReview.

- **Main Track:** The main track welcomes submissions of up to five pages, excluding references and supplementary materials, which have no page limit. We will welcome submissions that present work which is unpublished or currently under submission. We will also consider recently published (i.e., in 2025 or late 2024) work in venues other than ICLR 2025 and NeurIPS 2024.
- **Tiny Papers Track:** In addition to our main track, we offer a "Tiny Papers Track" for shorter works (up to 2 pages). This track encourages the presentation of works-in-progress and intermediate research milestones. We particularly encourage submissions from underrepresented, under-resourced, and early-career researchers to share their experiences, gather feedback, and foster collaboration.

B. Reviewing

All submitted papers will be reviewed by at least 3 experienced reviewers.

TABLE I: Tentative workshop schedule. All talks include a Q&A session.

Time (GMT+1)	**Session**	**Speaker(s)**
9:00 am - 9:05 am	Opening remarks	Organizers
9:05 am - 9:40 am	Mentoring Session	
9:40 am - 10:15 am	Invited talk 1	Atlas Wang
10:15 am - 10:35 am	Oral presentation 1	
10:35 am - 10:55 am	Oral presentation 2	
10:55 am - 11:00 am	Short break	
11:00 am - 12:00 pm	Poster session 1	
12:00 pm - 13:00 pm	Lunch break / Mentor-Mentee Lunch	
1:00 pm - 1:35 pm	Invited talk 2	Beidi Chen
1:35 pm - 2:10 pm	Invited talk 3	Sara Hooker
2:10 pm - 3:10 pm	Panel discussion	Panelists
3:10 pm - 3:15 pm	Short break	
3:15 pm - 4:15 pm	Poster session 2	
4:15 pm - 4:50 pm	Invited talk 4	Azalia Mirhoseini
4:50 pm - 5:25 pm	Invited talk 5	Dan Alistarh
5:25 pm - 5:45 pm	Oral presentation 3	
5:45 pm - 6:05 pm	Oral presentation 4	
6:05pm - 6:35pm	Breakout Session	
6:35 pm - 6:45 pm	Closing remarks and Award Ceremony	Organizers

a) Conflict of interest: We will evaluate potential conflicts of interest using OpenReview, and request that reviewers disclose any possible conflicts (e.g. papers from the same organization or previous supervisory relationships.). Upon receiving such notifications, the workshop program chairs will reassign the affected papers accordingly.

b) Selection of Oral Presentations: During the paper review process, each reviewer will have the opportunity to nominate a work for oral presentation. This nomination will apply to both the Main and Tiny Papers tracks to ensure that promising early research results are not overlooked. Once all reviews are submitted, the nominated papers will be made available to all non-conflicted reviewers, who will then vote on their preferences by ranking the nominated works.

c) Best Student Paper Award: This award will be given to the paper authored or co-authored by a student as an accolade to recognize their research potential and give them more visibility at the workshop/conference.

C. Tentative Schedule

SLLM will be a full-day in-person workshop. We plan to break the workshop into morning and afternoon sessions. Our schedule includes six invited talks, four short oral presentations, a panel discussion, two poster sessions, a breakout session, and a mentoring session.

a) Invited Talks: The invited talks will feature 35-minute presentations from well-know leaders in the field of sparse LLMs (see Section IV for a list). We will provide Q&A sessions for each scheduled talk, taking the questions both from live audiences and online chats.

b) Oral Presentations: The approximately 20-minute oral presentations (with Q&A) will showcase the best papers accepted to the workshop and will appear in the same session block as invited talks.

c) Poster Session: The poster sessions will offer researchers a chance to present their accepted work from both the main and short paper tracks, giving participants the opportunity to explore current topics in the field and engage directly with the authors. We plan to hold two sessions—one in the morning and one in the afternoon—interspersed with other workshop activities.

d) Panel: The panel discussion will center on evaluating *where we are, the limitations, and opportunities in improving LLM inference efficiency.* We will coordinate with the panelists in advance to discuss key questions and perspectives, but the primary focus will be on questions from the audience, encouraging active participant engagement. We hope that some of the ideas raised during this session will carry over into the breakout discussions later in the program.

e) Breakout Session: The breakout session will provide an interactive environment for participants to engage in small-group discussions. In advance, we will organize several groups based on key topics covered in the program and offer suggested questions to help initiate conversations. The goal of the breakout session is to create a forum where participants can share their insights and feedback on the workshop topics, engage in meaningful dialogue, and explore potential new research questions, bringing together scientists with shared interests in future directions.

f) Mentorship session: We aim to support young researchers, especially those from underrepresented groups in academia and research, by connecting them with senior researchers through a mentorship program. The morning mentorship session will serve as a kickoff meeting, introducing mentors and mentees to one another. This session will focus on providing guidance for navigating challenges in research and academia, fostering research collaborations, and offering advice on publishing, securing funding, and building professional relationships. Mentorship will also continue during lunch and asynchronously throughout the event.We will select senior researchers from diverse background to serve as mentors. The target audience for the mentoring session will be undergraduate, graduate and post-docs, with a priority focus of underrepresented groups.

We present the tentative workshop schedule in Table I. We have deliberately interwoven talk-focused sessions with activities that actively engaged the audience, like panels, poster sessions and breakout-discussions, offering a variety of events throughout both the morning and afternoon programs. Additionally, we aim to support early-stage researchers by offering mentorship programs and recognizing outstanding work through best student awards (see also Section III). Our goal is to engage the audience through these diverse formats, creating a space for participants to develop ideas, laying the groundwork for future research and lasting collaborations in the field.

D. Anticipated audience size

We anticipate an audience of 100–200 for the workshop, as topics such as training efficiency and inference efficiency are of interest to both industry and academic audiences. This estimate is based on our survey of attendance of the most related previous workshops.

III. Diversity and Inclusivity Commitment

Diversity is a fundamental value of our workshop, and we are deeply committed to fostering inclusivity in AI research. Through our exploration of various sparsity techniques in LLMs, we aim to enhance the accessibility of these models, reducing barriers to entry. This enables broader participation from smaller organizations, independent researchers, and practitioners from diverse backgrounds, empowering them to engage in state-of-the-art AI advancements. The compute-intensive nature of LLMs hinders accessibility to LLMs for people outside G7 countries. For example, as of June 2023, the G7 countries and China alone have 81% of the world's top 500 supercomputers, whereas Africa has only 0.2% of the computing resources. Even in G7 countries, universities, hospitals and startups do not have easy access to large-scale computing necessary for using LLMs. This disparity in access to computing resources creates a barrier to doing research and using LLMs. Our workshop will help bring researchers together and facilitate discussion to make LLMs more efficient and more easily accessible to people/researchers outside G7 countries.

a) Diversity in the Organizing Committee: Our organizing committee and roster of invited speakers showcase a wide array of diversity. We have intentionally selected speakers from various **affiliations**, encompassing both the academic and industrial sectors. Our speakers hail from diverse **geographical locations** including the US, UK, Europe, and Asia, and their **seniority** ranges from PhD students and postdoctoral researchers to research scientists, assistant professors, and full professors. Additionally, our organizing team is **gender** diverse, and represents a variety of **cultural backgrounds** from the US, Asia, and Europe. Notably, our team includes Mohammed Adnan, Tianlong Chen, and Shiwei Liu who have not organized before, bringing fresh perspectives to our workshop organization.

b) Early-Career, Independent, and Student Researchers: We understand that conducting ML research and running experiments often requires access to computing resources, which may not be readily available to everyone—especially students working independently without the support of a faculty advisor or PI. To address this, we will have a **Tiny Papers track**, considering that not everyone will have the resources to run experiments on large benchmarks. The Tiny Papers track will allow young researchers to present their

preliminary findings and ideas on a small dataset to receive feedback from experts in the workshop and potentially find mentors to further develop their research ideas.

To encourage the participation of emerging scholars, we are offering a **Best Student Paper Award** for student submissions, promoting excellence and innovation among junior researchers. We will also provide **mentorship** programs that pair early-career researchers and participants from underrepresented groups with experienced faculty. We are seeking sponsors to provide **travel funding** for students, as well as individuals from underrepresented and marginalized groups.

IV. Confirmed Invited Speakers

The following speakers have **all confirmed participation**.

(listed alphabetically by last name)

Dan Alistarh (IST Austria) is a professor in Computer Science at the Institute of Science and Technology (IST) Austria. His research focuses on distributed optimization and concurrent data structures, spanning from purely theoretical results to practical implementations. Recently, he has made pioneering contributions to efficient machine learning systems, including efficient federated learning methods like QSGD [55], and pruning [17] and quantization [56] of foundation models for efficient inference.

Beidi Chen (Carnegie Mellon University) is an Assistant Professor of Electrical and Computer Engineering at Carnegie Mellon University. She was previously a visiting researcher at Meta/Facebook AI Research (FAIR), postdoc researcher at Stanford working with Dr. Chris Ré, and received her Ph.D. in Computer Science from Rice University under the supervision of Dr. Anshumali Shrivastava in 2020. Beidi's research focuses on large-scale machine learning. Specifically, she designs and optimizes randomized algorithms (algorithm-hardware co-design) to accelerate large machine learning systems for real-world problems.

Sara Hooker (Cohere) is a VP of Research at Cohere and leads Cohere For AI, which is a non-profit organization to promote collaborative work and mentor young researchers. Previously, she was a research scientist at Google Brain. Sara has done seminal work in sparse neural networks, model efficiency, LLMs and algorithmic bias and fairness in machine learning. Recently, she was listed as one of top 100 influential people in AI by TIME.

Azalia Mirhoseini (Stanford/Google Deepmind) is an Assistant Professor of computer science at Stanford University, where she leads the Scaling Intelligence Lab. She also spends time at Google Deepmind, where she works as a Senior Staff Scientist. Her past work includes Mixture-of-Experts (MoE) neural architectures, now commonly used in frontier generative AI models, and reinforcement learning for chip floorplanning, a pioneering work in AI for chip design which has been used to design advanced AI accelerators and embedded chips.

Vithu Thangarasa (Cerebras Systems) Vithu is currently a Senior Machine Learning Research Scientist at Cerebras Systems focused on the intersection of systems and machine learning, with a strong interest in software/hardware co-design. Vithu's research focus is on efficient deep learning research at Cerebras, exploring unstructured sparse training and inference for LLMs, as well as distillation and speculative decoding [57], [58]. His work emphasizes advancing sparsity research to improve training and inference efficiency, demonstrating his commitment to innovation and passion for addressing complex challenges in machine learning. Vithu has in the past contributed to innovative projects at Uber AI Labs, and Tesla.

Atlas Wang (University of Texas at Austin/XTX Markets) 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 was its inaugural Program Chair.

Fuzhao Xue (National University of Singapore) is completing his PhD at NUS under Prof. Yang You. He has worked at NVIDIA with Jim Fan and Google Deepmind with Yi Tay and Mostafa Dehghani. Fuzhao's research focuses on Transformer Scaling, Adaptive and Conditional Computation, and Machine Learning

Systems. He has contributed to projects like Sequence Parallelism, AdaTape, Token-Crisis, OpenMoE and VILA. He is a recipient of the Google PhD Fellowship.

V. Advertising the Workshop

We will have a strong social media presence by creating a website, Twitter handle, and Discord server for the workshop. One organizer will be assigned as publicity chair, with a social media presence planned on common platforms used in the community. Additionally, we will be e-mailing out to popular mailing lists and newsgroups for researchers in machine learning the workshop call for papers, along with following up with the organizers' significant professional networks to personally invite researchers in academia and industry to participate. Finally, we will be encouraging early-career and aspiring researchers to participate, e.g. in our tiny papers call, by advertising the workshop within popular open research communities, such as ML Collective and Cohere 4 AI. Organizers of this workshop belong to different institutes and will also promote the workshop at their institute to help reach it to a wider audience. We will also ask invited speakers to advertise the workshop at their institutes/slack channels.

VI. Virtual access to workshop materials and outcome

The talks will be livestreamed and recorded. For each presentation and panel discussion, we plan to take questions from both the on-site audience and live chats, supporting the contribution of participants who could not attend the event in person. Moreover, a website will be set up and maintained, providing key information about the workshops, news updates, and documentation of the materials. In particular, accepted papers from both the main- and tiny-papers tracks will be made available online, along with a dedicated page for posters from the planned poster sessions. Additionally, we will publish the key ideas and concepts developed during the breakout sessions to allow participants to further work on those problems.

We will offer participants the chance to provide anonymous feedback on the content, activities, and format of the workshops, as well as their perceptions of the impact of the mentoring program and breakout sessions, through a feedback form hosted virtually. We will also encourage the participants to stay in touch, by joining the workshop's Discord channel, which we plan to keep active beyond the duration of the conference. In addition, we also maintain a sparsity Google e-mailing group for connecting people working in the topics discussed in the event.

VII. Related Previous Workshops

We will note that there has been a lack of workshops directly on the focus of the proposed workshop, i.e. inference in LLMs, and bringing together researchers across such a diverse set of research areas towards this. The following related workshops at ICML/ICLR/NeurIPS have been held:

- **ICLR 2024**: Reliable and Responsible FMs (focused on responsible AI), Workshop on practical ML for limited/low resource settings (focused on deployment of ML systems in compute constrained setting), Mathematical and Empirical Understanding of FMs (focused on understanding FMs).
- **ICML 2024**: Efficient Systems for FMs, Accessible and Efficient Foundation Models for Biological Discovery, Workshop on Advancing Neural Network Training (WANT).
- **NeurIPS 2024**: Workshop on Efficient Natural Language and Speech Processing (ENLSP), Scalable Continual Learning for Lifelong Foundational Models, Machine Learning and Compression Workshop
- **ICLR 2023**: The 3rd Sparsity in Neural Networks Workshop (focused on sparse training).
- **ICML 2022**: Workshop on Dynamic Neural Networks, The 2nd Sparsity in Neural Networks Workshop.

Making large language models (LLMs) more efficient has proved to be a complex challenge, requiring ideas and inputs from diverse areas, such as quantization, sparse training, model compression/pruning, and a deeper theoretical understanding of LLMs. While previous workshops have focussed a subset of these areas, our workshop will bring together, for the first time, researchers and practitioners from all these research areas and across both software and hardware-focused industry and academic groups to discuss potential research directions for making LLMs more efficient. For example, the **Sparsity in Neural Networks Workshop** focused on sparse training and pruning neural networks but did not focus on LLMs and only focused on weight/parameter sparsity. **WANT** specifically focused on scaling neural network training, whereas our workshop will focus on algorithmic and hardware innovation for faster inference as well. **Efficient Systems for FMs** had a similar high-level mandate to ours, however, our workshop will focus on additional

research topics, such as activation sparsity for inference and sparsity for interpretability. **Scalable Continual Learning for Lifelong Foundational Models** only focused on scalable approaches for life-long learning of foundational models. **Workshop on practical ML for limited/low resource settings** focused on training with limited data and model compression but did not focus on LLMs explicitly and other areas of efficiency. Our workshop, thus, will promote the fusion of ideas from various research topics, which is much required for tackling the efficiency problems in LLMs. Our workshop will allow, for the first time, a platform for researchers working on different aspects of efficiency — both algorithmic and hardware — to discuss and brainstorm new research directions and ideas.

VIII. Organizer Biosketches

(listed alphabetically by last name)

Points of Contacts: Yani Ioannou and Ashwinee Panda.

Tianlong Chen [Website, Google Scholar, Email: tianlong@cs.unc.edu] Tianlong received his Ph.D. degree in Electrical and Computer Engineering from the University of Texas at Austin, TX, USA, in 2023. He is currently an Assistant Professor of Computer Science at The University of North Carolina at Chapel Hill. He was previously a Postdoctoral Researcher at CSAIL@MIT, BMI@Harvard, and Broad Institute of MIT & Harvard. 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, and AdvML Rising Star Award. He has co-organized several tutorials in ICASSP'24, AAAI'24, and ICML'24.

Utku Evci [Website, Google Scholar, Email: evcu@google.com] is a Research Scientist in the Google DeepMind team in Montreal and studies efficient training and adaptation of neural networks. He participated in the Google AI Residency Program during 2018-2020 after completing his M.Sc. degree in Computer Science at NYU Courant. Utku co-led the organization of the first two Sparsity in Neural Network Workshops (sparseneural.net) and also co-leads the sparsity research group at Google DeepMind.

Yani Ioannou [Website, Google Scholar, Email: yani.ioannou@ucalgary.ca] is a Schulich Research Chair and Assistant Professor in the Department of Electrical and Software Engineering, in the Schulich School of Engineering at the University of Calgary, in Alberta, Canada. Yani leads the Calgary Machine Learning Lab, with a research focus on improving the efficiency and fairness/bias of deep learning models and efficient deep learning methods. Previously, he was a Postdoctoral Research Fellow at the Vector Institute, and Visiting Researcher at Google Brain Toronto. Yani was awarded a Microsoft Research PhD Scholarship, and completed his PhD at the University of Cambridge in 2018 under the supervision of Roberto Cipolla and Antonio Criminisi, while also a visiting student at Microsoft Research Cambridge.

Berivan Isik [Website, Google Scholar, Email: berivan@google.com] is a research scientist at Google, working on efficient and trustworthy AI. Her current interests are efficient training/finetuning of large models, pretraining data valuation and scaling laws for LLMs, differential privacy, and unlearning. She completed her PhD at Stanford University, advised by Tsachy Weissman and Sanmi Koyejo, where she was affiliated with the SAIL and StatsML groups. Her research was supported by Stanford Graduate Fellowship (2019-2023), Google Ph.D. Fellowship (2023-2026), and a Meta research grant. She has co-organized various workshops in the past, including TF2M-ICML'24, DMLR-ICML'24, Neural Compression Workshop at ICML'23, ITR3-ICML'21, and WiML-ICML'21.

Shiwei Liu [Website, Google Scholar, Email: shiwei.liu@maths.ox.ac.uk] is a Royal Society Newton International Fellow at 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 of large-foundation models, robustness and trustworthiness, and generative AI. Dr. Liu has received two 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 co-organized several tutorials in ICASSP'24, IJCAI'23, and ECML-PKDD'22, as well as the Edge-Device LLM Challenge and Workshop in NeurIPS'24.

Adnan Mohammed [Website, Google Scholar, Email: adnan.ahmad@ucalgary.ca] is a PhD student at the University of Calgary and Vector Institute, working under the supervision of Dr. Yani Ioannou (University of Calgary) and Dr. Rahul Krishnan (University of Toronto/Vector Institute). His research interests include efficient machine learning, understanding loss-landscape properties of neural networks and applications of Neural Tangent Kernels. His research is supported by the NSERC Doctoral Fellowship, Borealis AI Research Fellowship, and Digital Research Alliance of Canada. He received his MS and undergraduate degrees from the University of Waterloo and the Indian Institute of Technology (IIT) Guwahati, respectively.

Aleksandra Nowak [Website, Google Scholar, Email: aleksandrairena.nowak@doctoral.uj.edu.pl] is a PhD student at the Jagiellonian University in Cracow, working within the Group of Machine Learning Research led by prof. Jacek Tabor. Her research interests include the analysis and development of sparse neural network architectures and efficient adaptation techniques. In the past years, she has co-organized the EEML 2020 Summer School, the MLSS^N 2022 Summer School, and the MLinPL 2021 Conference. She is also a member of the MLinPL Association, a non-profit organization devoted to fostering the machine learning community in Poland.

Ashwinee Panda [Website, Google Scholar, Email: ashwinee@umd.edu] is a postdoctoral fellow at the NSF Institute for Trustworthy AI in Law and Society (TRAILS) working with Prof. Tom Goldstein on LLM pretraining. Ashwinee received his PhD from Princeton, where he was advised by Prof. Prateek Mittal, working on "Unlocking Trustworthy Machine Learning with Sparsity".

IX. References

[1] J. Kaplan, S. McCandlish, T. Henighan, *et al.*, *Scaling laws for neural language models*, 2020. arXiv: 2001.08361 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2001.08361.

[2] J. Hoffmann, S. Borgeaud, A. Mensch, *et al.*, *Training compute-optimal large language models*, 2022. arXiv: 2203.15556 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2203.15556.

[3] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention is all you need*, 2023. arXiv: 1706.03762 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1706.03762.

[4] G. Team, P. Georgiev, V. I. Lei, *et al.*, *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*, 2024. arXiv: 2403.05530 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2403.05530.

[5] xAI, *Grok-1*, 2024. [Online]. Available: https://github.com/xai-org/grok-1?tab=readme-ov-file.

[6] Databricks, *Dbrx*, 2024. [Online]. Available: https://www.databricks.com/blog/introducing-dbrx-new-state-art-open-llm.

[7] A. Q. Jiang, A. Sablayrolles, A. Roux, *et al.*, *Mixtral of experts*, 2024. arXiv: 2401.04088 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2401.04088.

[8] Snowflake, *Arctic*, 2024. [Online]. Available: https://www.snowflake.com/en/blog/arctic-open-efficient-foundation-language-models-snowflake/.

[9] DeepSeek-AI, A. Liu, B. Feng, *et al.*, *Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model*, 2024. arXiv: 2405.04434 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2405.04434.

[10] N. Shazeer, A. Mirhoseini, K. Maziarz, *et al.*, *Outrageously large neural networks: The sparsely-gated mixture-of-experts layer*, 2017. arXiv: 1701.06538 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1701.06538.

[11] A. Clark, D. de las Casas, A. Guy, *et al.*, *Unified scaling laws for routed language models*, 2022. arXiv: 2202.01169 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2202.01169.

[12] N. Du, Y. Huang, A. M. Dai, *et al.*, "Glam: Efficient scaling of language models with mixture-of-experts," in *International Conference on Machine Learning*, PMLR, 2022, pp. 5547–5569.

[13] D. Lepikhin, H. Lee, Y. Xu, *et al.*, "Gshard: Scaling giant models with conditional computation and automatic sharding," *arXiv preprint arXiv:2006.16668*, 2020.

[14] W. Fedus, B. Zoph, and N. Shazeer, *Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity*, 2022. arXiv: 2101.03961 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2101.03961.

[15] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive Mixtures of Local Experts," *Neural Computation*, vol. 3, no. 1, pp. 79–87, Mar. 1991, ISSN: 0899-7667. DOI: 10.1162/neco.1991.3.1.79. eprint: https://direct.mit.edu/neco/article-pdf/3/1/79/812104/neco.1991.3.1.79.pdf. [Online]. Available: https://doi.org/10.1162/neco.1991.3.1.79.

[16] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the em algorithm," in *ICANN '94*, M. Marinaro and P. G. Morasso, Eds., London: Springer London, 1994, pp. 479–486, ISBN: 978-1-4471-2097-1.

[17] E. Frantar and D. Alistarh, "Sparsegpt: Massive language models can be accurately pruned in one-shot," in *International Conference on Machine Learning*, PMLR, 2023, pp. 10 323–10 337.

[18] M. Sun, Z. Liu, A. Bair, and J. Z. Kolter, "A simple and effective pruning approach for large language models," in *The Twelfth International Conference on Learning Representations*, 2024.

[19] A. Jaiswal, Z. Gan, X. Du, B. Zhang, Z. Wang, and Y. Yang, "Compressing llms: The truth is rarely pure and never simple," in *ICLR*, 2024. [Online]. Available: https://arxiv.org/abs/2310.01382.

[20] S. Ma, H. Wang, L. Ma, *et al.*, "The era of 1-bit llms: All large language models are in 1.58 bits," *ArXiv*, vol. abs/2402.17764, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268041246.

[21] S. Ashkboos, A. Mohtashami, M. L. Croci, *et al.*, "Quarot: Outlier-free 4-bit inference in rotated llms," *ArXiv*, vol. abs/2404.00456, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268819214.

[22] S. T. Sreenivas, S. Muralidharan, R. Joshi, *et al.*, "Llm pruning and distillation in practice: The minitron approach," *ArXiv*, vol. abs/2408.11796, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:271915771.

[23] Y. Song, H. Xie, Z. Zhang, *et al.*, "Turbo sparse: Achieving llm sota performance with minimal activated parameters," *arXiv preprint arXiv:2406.05955*, 2024.

[24] I. Mirzadeh, K. Alizadeh, S. Mehta, *et al.*, *Relu strikes back: Exploiting activation sparsity in large language models*, 2023. arXiv: 2310.04564 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2310.04564.

[25] Z. Liu, J. Wang, T. Dao, *et al.*, "Deja vu: Contextual sparsity for efficient llms at inference time," in *International Conference on Machine Learning*, PMLR, 2023, pp. 22 137–22 176.

[26] H. Wang, S. Ma, R. Wang, and F. Wei, "Q-sparse: All large language models can be fully sparsely-activated," *arXiv preprint arXiv:2407.10969*, 2024.

[27] T. Fu, H. Huang, X. Ning, *et al.*, *Moa: Mixture of sparse attention for automatic large language model compression*, 2024. arXiv: 2406.14909 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2406.14909.

[28] G. Xiao, Y. Tian, B. Chen, S. Han, and M. Lewis, "Efficient streaming language models with attention sinks," *arXiv preprint arXiv:2309.17453*, 2023.

[29] Z. Zhang, Y. Sheng, T. Zhou, *et al.*, "H2o: Heavy-hitter oracle for efficient generative inference of large language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[30] Y. Sheng, L. Zheng, B. Yuan, *et al.*, "Flexgen: High-throughput generative inference of large language models with a single gpu," in *International Conference on Machine Learning*, PMLR, 2023, pp. 31 094–31 116.

[31] Y. Zhao, C.-Y. Lin, K. Zhu, *et al.*, "Atom: Low-bit quantization for efficient and accurate llm serving," *Proceedings of Machine Learning and Systems*, vol. 6, pp. 196–209, 2024.

[32] Z. Liu, J. Yuan, H. Jin, *et al.*, "Kivi: A tuning-free asymmetric 2bit quantization for kv cache," *arXiv preprint arXiv:2402.02750*, 2024.

[33] O. Lieber, B. Lenz, H. Bata, *et al.*, *Jamba: A hybrid transformer-mamba language model*, 2024. arXiv: 2403.19887 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2403.19887.

[34] H. Jiang, Q. Wu, C.-Y. Lin, Y. Yang, and L. Qiu, "Llmlingua: Compressing prompts for accelerated inference of large language models," *arXiv preprint arXiv:2310.05736*, 2023.

[35] Z. Pan, Q. Wu, H. Jiang, *et al.*, "Llmlingua-2: Data distillation for efficient and faithful task-agnostic prompt compression," *arXiv preprint arXiv:2403.12968*, 2024.

[36] Y.-N. Chuang, T. Xing, C.-Y. Chang, Z. Liu, X. Chen, and X. Hu, "Learning to compress prompt in natural language formats," *arXiv preprint arXiv:2402.18700*, 2024.

[37] H. Cunningham, A. Ewart, L. Riggs, R. Huben, and L. Sharkey, "Sparse autoencoders find highly interpretable features in language models," *arXiv preprint arXiv:2309.08600*, 2023.

[38] S. Marks, C. Rager, E. J. Michaud, Y. Belinkov, D. Bau, and A. Mueller, "Sparse feature circuits: Discovering and editing interpretable causal graphs in language models," *arXiv preprint arXiv:2403.19647*, 2024.

[39] L. Gao, T. D. la Tour, H. Tillman, *et al.*, "Scaling and evaluating sparse autoencoders," *arXiv preprint arXiv:2406.04093*, 2024.

[40] S. Rajamanoharan, A. Conmy, L. Smith, *et al.*, "Improving dictionary learning with gated sparse autoencoders," *arXiv preprint arXiv:2404.16014*, 2024.

[41] S. Rajamanoharan, T. Lieberum, N. Sonnerat, *et al.*, *Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders*, 2024. arXiv: 2407.14435 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2407.14435.

[42] NVIDIA, *Structured sparsity in the nvidia ampere architecture and applications in search engines*, 2023. [Online]. Available: https://developer.nvidia.com/blog/structured-sparsity-in-the-nvidia-ampere-architecture-and-applications-in-search-engines/.

[43] Pytorch, 2024. [Online]. Available: https://pytorch.org/blog/accelerating-neural-network-training/.

[44] Y. N. Wu, P.-A. Tsai, S. Muralidharan, A. Parashar, V. Sze, and J. Emer, "Highlight: Efficient and flexible dnn acceleration with hierarchical structured sparsity," in *56th Annual IEEE/ACM International Symposium on Microarchitecture*, ser. MICRO '23, vol. 22, ACM, Oct. 2023, pp. 1106–1120. DOI: 10.1145/3613424.3623786. [Online]. Available: http://dx.doi.org/10.1145/3613424.3623786.

[45] Z. Y. Xue, Y. N. Wu, J. S. Emer, and V. Sze, "Tailors: Accelerating sparse tensor algebra by overbooking buffer capacity," in *56th Annual IEEE/ACM International Symposium on Microarchitecture*, ser. MICRO '23, ACM, Oct. 2023, pp. 1347–1363. DOI: 10.1145/3613424.3623793. [Online]. Available: http://dx.doi.org/10.1145/3613424.3623793.

[46] J.-P. Fricker, "The cerebras cs-2: Designing an ai accelerator around the world's largest 2.6 trillion transistor chip," in *Proceedings of the 2022 International Symposium on Physical Design*, 2022, pp. 71–71.

[47] N. Ding, Y. Qin, G. Yang, *et al.*, "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature Machine Intelligence*, vol. 5, no. 3, pp. 220–235, 2023.

[48] E. J. Hu, Y. Shen, P. Wallis, *et al.*, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[49] N. Houlsby, A. Giurgiu, S. Jastrzebski, *et al.*, "Parameter-efficient transfer learning for nlp," in *International conference on machine learning*, PMLR, 2019, pp. 2790–2799.

[50] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," *arXiv preprint arXiv:2104.08691*, 2021.

[51] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," *arXiv preprint arXiv:2101.00190*, 2021.

[52] E. B. Zaken, S. Ravfogel, and Y. Goldberg, "Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models," *arXiv preprint arXiv:2106.10199*, 2021.

[53] J. O. Zhang, A. Sax, A. Zamir, L. Guibas, and J. Malik, "Side-tuning: A baseline for network adaptation via additive side networks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, Springer, 2020, pp. 698–714.

[54] Y.-L. Sung, J. Cho, and M. Bansal, "Lst: Ladder side-tuning for parameter and memory efficient transfer learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 12 991–13 005, 2022.

[55] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "Qsgd: Communication-efficient sgd via gradient quantization and encoding," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6c340f25839e6acdc73414517203f5f0-Paper.pdf.

[56] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, "OPTQ: Accurate quantization for generative pre-trained transformers," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=tcbBPnfwxS.

[57] V. Thangarasa, S. Saxena, A. Gupta, and S. Lie, "Sparse-IFT: Sparse iso-FLOP transformations for maximizing training efficiency," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 235, PMLR, 2024. [Online]. Available: https://proceedings.mlr.press/v235/thangarasa24a.html.

[58] V. Thangarasa, G. Venkatesh, N. Sinnadurai, and S. Lie, *Self-data distillation for recovering quality in pruned large language models*, 2024. arXiv: 2410.09982 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2410.09982.