Edge-LLMs : Edge-Device Large Language Model Competition

Shiwei Liu[*] Kai Han[†] Adriana Fernandez-Lopez[‡] Ajay Jaiswal[§]
Zahra Atashgahi[¶] Boqian Wu[¶‖] Rebekka Burkholz[**] Edoardo M. Ponti[††]
Cong Hao[‡‡] Olga Saukh Lu Yin Tianjin Huang Andreas Zinonos
Jared Tanner[*] Yunhe Wang[†]

https://edge-llms-challenge.github.io/edge-llm-challenge.github.io/
Email: edgellmschallenge@gmail.com

Abstract

The Edge-Device Large Language Model Competition seeks to explore the capabilities and potential of large language models (LLMs) deployed directly on edge devices. The incredible capacity of LLMs makes it extremely tantalizing to be applied to practical edge devices to enable wide applications of LLMs in various disciplines. However, the massive size of LLMs poses significant challenges for edge devices where the computing resources and memory are strictly limited. For instance, deploying a small-scale 10B LLM could require up to 20GB of main memory (DRAM) even after adopting INT8 quantization, which unfortunately has exceeded the memory of most commodity smartphones. Besides, the high energy consumption of LLMs will drain smartphones' battery quickly. To facilitate applications of LLMs in a wide range of practical scenarios, we propose this timely competition to encourage practitioners in both academia and industry to come up with effective solutions for this pressing need. By challenging participants to develop efficient and optimized models that can run on resource-constrained edge devices, the competition aims to address critical economic and environmental issues related to LLMs, foster interdisciplinary research collaborations, and enhance the privacy and security of AI systems.

Keywords

Deep Learning, Large Language Models, Model Compression, Edge Computing, Edge LLM

[*]University of Oxford, UK
[†]Huawei Noah's Ark Lab, China
[‡]Meta AI, UK
[§]University of Texas at Austin, USA
[¶]University of Twente, the Netherlands
[‖]University of Luxembourg, Luxembourg
[**]Helmholtz Center CISPA, Germany
[††]University of Edinburgh, UK
[‡‡]Georgia Institute of Technology, USA
Graz University of Technology, Austria
University of Surrey, UK
University of Exeter, UK
Imperial College London, UK

Preprint. Under review.

1 Competition description

1.1 Background and impact

The rapid advancements in artificial intelligence, particularly in the domain of natural language processing (NLP) driven by Large Language Models (LLMs), have underscored the transformative potential of these colossal models in shaping modern modes of work and communication. Consequently, the prospect of integrating the expansive knowledge capacity of LLMs into edge devices, including smartphones, IoT devices, and in-car systems, holds significant promise within contemporary computing ecosystems. However, the massive size of LLMs presents formidable challenges for edge devices, which typically operate with constrained resources. For instance, deploying a modest 10B LLM can demand up to 20GB of main memory (DRAM), even after employing INT8 quantization techniques, a capacity that surpasses the available memory in most commodity smartphones [6]. For example, the HUAWAI P60 boasts an 8GB DRAM capacity, and the iPhone 15 offers a 6GB DRAM. Given that DRAM is shared among the operating system and other applications, mobile apps must not exceed 10% of the DRAM allocation [12]. Additionally, the energy consumption of LLMs presents a big barrier, with a fully charged smartphone, boasting approximately 50 kJ of energy, capable of sustaining a 7B LLM conversation for less than 2 hours at a rate of 10 tokens per second [11].

To this end, we propose this challenge aiming to push the boundaries of what these powerful LLMs can achieve on edge devices in terms of performance, efficiency, and versatility. Specifically, our goal is to deploy LLMs on edge devices in resource-bounded scenarios and address the following major issues:

- **Vast amounts of memory requirements:** LLM inference typically requires a significant amount of memory, a major bottleneck for off-the-shelf smartphones. Even high-end smartphones with 8GB are not sufficient for a sophisticated LLM.
- **Prohibitive energy consumption:** Moreover, the significant energy consumption during LLM inference presents a formidable challenge to the battery life of smartphones.
- **Large performance loss:** State-of-the-art LLMs might require unusually high compression ratios to fit into the memory of edge devices. Such extremely high compression ratios will significantly challenge the efficacy of existing model compression techniques, where we usually see a substantial performance drop after certain high levels of compression ratios. Therefore, achieving such high compression ratios while maintaining acceptable performance is very challenging.
- **Lack of offline functionality:** Most off-the-shelf LLMs necessitate internet connections, which limits their usage in remote areas or in situations where internet access may be intermittent.

Scope: This competition targets researchers, practitioners, and industry professionals from a wide range of fields, who are interested in co-designing systems, hardware, and algorithms to enable high-performing LLMs on edge devices, unlocking new possibilities for various industries and use cases. The competition is particularly relevant to the NeurIPS community. Topics related to LLMs, such as model architecture design, pre-training techniques, fine-tuning strategies, interpretability, and applications in various domains, are often featured prominently in NeurIPS presentations, workshops, and tutorials. Given the prevalence of LLMs and the challenges associated with deploying them on resource-limited yet ubiquitous edge devices, make NeurIPS an ideal platform for us to organize this competition. We anticipate that over 100 teams will participate in the competition, and we expect a large audience of around 400 people to attend the workshop.

Anticipated Impact: Our competition is expected to have significant impacts in terms of science, economics, and society. In terms of science, focusing on applying LLMs on edge devices can inspire the creation of innovative algorithms and system optimizations to bring powerful LLMs to these resource-constrained platforms. By enabling training and deployment of LLMs with academic-level computing resources, the competition will encourage more university researchers to study LLMs, fostering interdisciplinary collaborations across fields like linguistics, cognitive science, education, healthcare, and medicine. Economically, solutions that enable running LLMs on edge devices will reduce the energy costs of training and inference, leading to savings in network bandwidth, cloud infrastructure usage fees, and data transmission expenses. This can have a significant impact on the

affordability and accessibility of LLM-based applications. Socially, competitions focused on applying LLMs on edge devices can address pressing societal challenges such as digital inclusion, privacy, and environmental sustainability. Participants may develop solutions that democratize access to advanced language processing capabilities, protect user privacy through on-device processing, and reduce the carbon footprint associated with centralized cloud infrastructure. By leveraging edge computing and LLMs, competition solutions have the potential to drive positive social change, promote equity, and address the needs of underrepresented communities.

1.2 Novelty

This is a brand-new competition, which focuses on compressing and deploying LLMs on edge devices, such as smartphones, to push the boundaries of what these edge-device LLMs can achieve in terms of performance, efficiency, and versatility. This challenge is unique and distinct from the conventional model compression techniques due to several factors.

The first challenge arises from vast memory requirements. For instance, a 175B LLM requires at least 320GB of memory to store in half-precision (FP16) format. However, edge devices, such as smartphones and IoT devices, typically have very limited memory, posing a significant challenge for deploying and running LLMs. Even the latest Huawei Mate 60 Pro, equipped with 12GB DRAM, falls far short of the memory needed to support powerful LLMs. Consequently, innovative model placement strategies must be developed to reduce memory and storage requirements for edge LLMs.

The second challenge is the extreme requirements for computing capabilities. GPT-3 175b entails approximately 3–4 seconds to analyse a 60-token sentence and generate a 20-token response, even when running with 8 A100 GPUs [3]. The latency on commodity smartphones with low computing resources is expected to be much worse, possibly extending to minutes. Edge LLMs often need to operate in real-time or near-real-time, imposing further restrictions on latency and inference speed. Therefore, addressing the challenges of edge LLMs involves optimizing not only model size but also computational efficiency and inference speed to ensure practical real-time communications.

Last but not least, performance. Enabling LLMs on edge devices requires achieving unusually high compression rates to reduce the model size of billions of parameters to sub-billion. Such extremely high compression ratios will significantly challenge the efficacy of existing model compression techniques, as it is well known to be very challenging to achieve such high compression ratios while maintaining reasonable performance [18, 9]. In this context, we pose the following timely and important question:

Can we scale down the size of state-of-the-art LLMs to ***sub-million parameters*** *to accommodate the memory constraints of edge devices like smartphones, while preserving the robust performance of billion-parameter LLMs?*

To incentivize research efforts aimed at addressing this crucial research question, we introduce the "Edge-Device Language Model Challenge", which aims to develop sub-million parameter LLMs capable of efficient execution on mobile devices. To our knowledge, no previous competitions have specifically targeted this goal. One similar challenge is the "NeurIPS 2023 Large Language Model Efficiency Challenge: 1LLM + 1GPU + 1Day" [link]. However, their primary objective is orthogonal to ours, focusing on optimizing the fine-tuning of LLMs with limited resources, specifically 1GPU + 1Day, rather than deploying LLMs on edge devices. Even with a constraint of one single GPU, their solutions remain within the scenarios with reasonable resources, i.e., NVIDIA A100 with 40GB memory, which is way better than what a commodity smartphone can provide. Furthermore, their focus does not extend to achieving real-time or near-real-time interaction with low latency, which is a crucial aspect of deploying LLMs on edge devices.

Besides, there was "Applied AI Challenge: Large Language Models (LLMs)" [link] aiming to improve Federal Government Services through the use of LLMs, whose goal is completely different from ours.

1.3 Data

This competition will not evaluate submissions based on the analysis of data. Our competition features two tracks: (1) post-training LLM compression for edge devices; and (2) training edge LLMs from scratch. We will allow the participants to use and only use the C4 dataset [13] and (Chinese)

Alpaca [15, 14] for both tracks. The C4 dataset is a colossal, cleaned version of Common Crawl's web crawl corpus, which is mainly intended to pre-train language models and word representations. Language models like MPT-7B and T5 are pre-trained with the C4 dataset. C4 is a large enough dataset that can make the competition interesting and draw conclusive and statistically significant results. The C4 dataset is public and can be accessed through HuggingFace [1]. Alpaca is a dataset of 52,000 instructions and demonstrations generated by OpenAI's text-davinci-003 engine. This instruction data can be used to conduct instruction-tuning for language models and make the language model follow instructions better [15]. The Chinese version of Alpaca can be found here [14].

To comprehensively evaluate the submissions, we meticulously collect a diverse range of *public* datasets, ensuring the inclusion of domain-specific and intricate knowledge to capture a wide spectrum of capabilities such as language comprehension, knowledge precision, logical deduction, mathematical problem-solving, programming proficiency, extended text analysis, and intelligent agent engagement. Please see Table 1 for more details.

1.4 Tasks and application scenarios

This competition will have the following tracks: (1) post-training LLM compression for edge devices; and (2) training edge LLMs from scratch.

For the first track, participating teams are requested to come up with their own compression approaches to compress three pre-trained LLMs separately, i.e., Phi-2, Llama3-7B, and Qwen-7B. Every single model in the submission is required to be run on a smartphone with 12 GB DRAM. Submissions will be ranked based on the averaged scores of three models on a subset of the OpenCompass benchmark [5], which performs an in-depth and holistic assessment of LLMs across multiple fundamental dimensions. Our second track requires training language models from scratch without leveraging any pre-trained LLMs. This track has no constraints on model architectures, training recipes, and training time, as long as the final models can be run on a 12GB smartphone. Participants can design their architectures or leverage existing LLM architectures to fulfill this task. However, we do have a constraint on the training data, i.e., only the C4 and Alpaca datasets can be used for training and fine-tuning, respectively. **Note that no quantization methods are allowed as the 8-bit or 4-bit quantization for LLMs is already a well-established technique. The participants are required to submit models in FP16 or FP32 format.**

The two tracks outlined above address urgent, real-world scenarios driven by the growing demand for on-device natural language processing capabilities, a pressing issue in the industry. For example, the deployment of robust pre-trained LLMs directly onto smartphones is vital for applications like virtual assistants, voice interfaces, language translation, and personalized content recommendations. It is a long-standing technique in the industry to compress pre-trained large models to small ones for practical deployment. However, the colossal size of state-of-the-art LLMs necessitates exceptionally high compression ratios—often over 100 times—to fit them onto smartphones. Achieving such compression ratios while maintaining the reasonable performance of LLMs presents a significant challenge [18, 9]. Conversely, designing new architectures that are more suitable to edge devices with smaller sizes can be an alternative way to achieve this goal. This motivation underpins our decision to feature the second track: training edge LLMs from scratch.

While both tracks pose significant challenges to our community, it is not impossible to solve them. We have seen research endeavors making progress in compressing LLMs [8, 17, 18], and designing new architectures for mobile devices [11], even though the performance is still far from satisfactory. Our challenge serves as timely motivation to inspire more researchers to tackle this urgent and critical problem.

1.5 Metrics

To comprehensively evaluate submissions, we employ a set of rigorously curated metrics, which include:

- **Performance Score**: One of our primary scoring metrics is the performance score on the selected evaluation task. Each submission will be ranked individually for each task based on the performance score, with the top 10 submissions per task receiving scores from 10 to 1. The final score of the submission is calculated as the sum across all evaluation tasks.

- **Memory Requirement:** The memory footprint during inference is a crucial metric for real-life edge devices. To qualify, the peak memory usage of all models must be less than 12GB.
- **Throughput**: The throughput of an LLM typically refers to the rate at which the model can process input data and generate output tokens. It is often measured in terms of tokens per second. Throughput is a critical metric for evaluating the efficiency and performance of LLMs, especially in real-time or near-real-time applications where the speed of processing is crucial. Achieving high throughput implies that the model can handle large volumes of data quickly, making it suitable for tasks requiring rapid language processing, such as live chatbots, real-time translation, or speech recognition systems. This value will be measured on a smartphone with 12GB DRAM.
- **Parameter count**: Submissions should also include the model size, expressed as the parameter count, to indicate the model's size. However, this metric is used for information only, not for ranking.

1.6 Baselines, code, and material provided

We will release the *"starting kit"* to provide a starting point for people who are interested in our challenge before June 25th, 2024. This starting kit will provide detailed clarifications on what a submission looks like exactly, and how it will be evaluated and submitted. The starting kit will include an end-to-end submission flow, exemplified with a simple baseline:

- Loading a large language model choosing from Phi-2, Llama3-8B, or Qwen-7B.
- Performing the basic L1 norm structured pruning on this model so that the compressed model can fit the smartphone.
- Evaluating the pruned model on the OpenCompass benchmark. We will provide an easy-to-tun pipeline for participants to evaluate their model on the subset of the OpenCompass benchmark.
- Deploying the resulting model on a smartphone platform and measuring its throughput. We will provide a tool that can help participants easily deploy their LLMs on the smartphone platform and measure throughput.

1.7 Website, tutorial and documentation

We have developed a tentative website for our competition, accessible through the following link: `https://edge-llms-challenge.github.io/edge-llm-challenge.github.io/`. This website will be self-contained and present all relevant information about the competition timeline and illustrate the necessary steps to participate. Additionally, we have included a FAQ/Tutorial section and a dedicated email address, i.e., `edgellmschallenge@gmail.com`, to facilitate communication with the organizers. We already have a robust code base to assess submission performance across various tasks. We anticipate that all content, including the starter kit, will be available online by May 25th, 2024.

2 Organizational aspects

2.1 Protocol

Our competition will consist of two tracks: (1) Post-training LLM Compression for Edge Devices; and (2) Training Edge LLMs from Scratch. Participants have the option to choose either track or both for their participation in the competition.

Track 1: Compression Challenge. In this track, participating teams are tasked with developing their own compression methods to compress three pre-trained LLMs individually: Phi-2, Llama-3-8B, and Qwen-7B. Each model submitted must be capable of running on a smartphone device with 12 GB DRAM. Each model will be evaluated on a subset of the OpenCompass benchmark [5], which comprehensively assesses LLMs across multiple fundamental dimensions. We will take the average score of three models as the final score of the submission for each task.

Track 2: Training from Scratch Challenge. In this track, participants are challenged to train language models from scratch without utilizing any pre-trained LLMs. There are no constraints on model architectures, training procedures, or duration, as long as the final models can run on a smartphone device with 12GB memory. Participants are free to design their architectures or utilize existing LLM architectures for this task. However, there is a restriction on the training data: only the C4 and Alpaca datasets are permitted for training and fine-tuning.

Note: Quantization methods are not allowed, as 8-bit or 4-bit quantization for LLMs is a well-established technique. Participants must submit models in FP16 or FP32 format.

Approved LLMs and dataset. To promote diverse participation, our approved pre-trained LLMs encompass three representative LLMs as our base models for Track 1, including Phi-2, Qwen-7B, and Llama3-8B. There is no architecture constraint for Track 2.

- **Llama3-8B** [16] is the latest version of Meta's large language model. Llama3-8B is pre-trained on over 15T tokens, which is seven times larger than that used for Llama2, and it includes four times more code. To prepare for upcoming multilingual use cases, over 5% of the Llama 3 pre-training dataset consists of high-quality non-English data that covers over 30 languages.
- **Phi-2** [4] is a 2.7 billion-parameter language model of Microsoft with a context length of 2048 tokens. It was scaled from the 1.3 billion parameter model, Phi-1.5, and trained with a mixture of data containing synthetic datasets specifically created to teach the model common sense reasoning and general knowledge. The training data contains 1.4T tokens.
- **Qwen-7B.** Tongyi Qianwen-7B (Qwen-7B) [7] is a 7 billion parameter scale model of the Tongyi Qianwen large model series developed by Alibaba Cloud. Qwen-7B is a LLM based on Transformer, which is trained on extremely large-scale pre-training data. The pre-training data types are diverse and cover a wide range of areas, including a large number of online texts, professional books, codes, etc.

We have selected C4 [13] and alpaca as the exclusive datasets permitted for this competition. We will allow the participants to use and only use the C4 dataset [13] for model compression and pre-training. In addition to C4, the Alpaca, and Chinese Alpaca datasets are allowed for supervised fine-tuning to ensure the quality of trained models. C4, derived from the Common Crawl dataset [2], offers several advantages. Firstly, it is freely available to the public and has been widely utilized in training LLMs, including T5 text-to-text Transformer models and MPT-7B. Secondly, it demonstrates performance comparable to popular LLM pre-training datasets like Pile [10]. Lastly, restricting the competition to the C4 dataset helps ensure fairness by eliminating potential advantages gained from closed-source datasets. Three models are expected to be submitted for Track 1 and only one model is expected to be submitted for Track 2.

Evaluation. The evaluation process in our competition will include two stages.

The first stage: the submitted models will be evaluated on a diverse subset of the OpenCompass benchmark, including CSQA, BIG-Bench Hard, GSM8K, LongBench, HumanEval, T-EVAL, CHID, along with a set of secret holdout tasks to avoid overfitting. A self-contained code base will be provided to facilitate participants in easily evaluating their models on seven diverse tasks via the OpenCompass benchmark.

- Each submission will be ranked individually for each task, with the top 10 submissions with the highest score per task receiving scores from 10 to 1. The score of each task is calculated as the average score of the three models for the first track. For the second track, the score is the score of the submitted model.
- Additionally, we will measure the models' throughput on a smartphone platform provided by the sponsor. Submissions will be ranked based on throughput. To emphasize the importance of inference speed, the score for the throughput task will be **doubled**, ranging from 20 to 2 for the top 10 submissions. To assist participants in improving the speed of their models, we will provide an easy-to-run pipeline to measure throughput on a GPU, where the throughput values are roughly scaled to those on the smartphone platform.
- We will measure the inference memory usage of all models. Submissions with models that exceed 12GB of memory usage for inference will be disqualified.

- The final rank of submissions will be determined by the sum of scores across all evaluation tasks including seven diverse tasks (100 scores in total) as shown in Table 1 (maximum 70 scores), one secret holdout task (maximum 10 scores), and the throughput measurement (maximum 20 scores). To be more specific, the final score of submission is $\sum_{i=1}^{n} s_i$, where s_i is the score of each evaluation task.

The top 15 teams with the highest scores will be displayed on the leaderboard. Participants are required to submit their source codes, evaluation log files, and models to the organizers via their own GitHub repositories, with models accessible through a valid Google Drive link.

Participating teams are encouraged to submit their models for a preliminary review (25th August) to identify potential bugs and format issues, ensuring that the final submissions are in the correct format. Each team can make three submissions to each track after the deadline of the preliminary review. The best-performing model will be used to rank the team on the leaderboard and selected for the final model evaluation.

The second stage: After the competition is closed on October 25th, 2024, we will contact the top 3 teams with the highest scores in both tracks, requesting that they submit all necessary code and data to reproduce their results. We will then replicate their entire process, to ensure it is fully repeatable and the final model can be run on a smartphone with 12 GB RAM with the same results. If the top-scoring model cannot be reproduced under these imposed conditions or can not fit the smartphone, we will move on to consider the next highest-scoring submission in the category, until a reproducible and high-performing submission is selected.

Our evaluation will leverage an extensive array of rigorously curated datasets across multiple fundamental dimensions: language comprehension, knowledge precision, logical deduction, mathematical problem-solving, programming proficiency, extended text analysis, and intelligent agent engagement. Details of the evaluation tasks we chose are shown below:

Table 1: Evaluation Dataset of EdgeLLM Competition.

Dataset	Dimension	Source
CommonsenseQA	Knowledge	https://www.tau-nlp.sites.tau.ac.il/commonsenseqa
BIG-Bench Hard	Reasoning	https://github.com/suzgunmirac/BIG-Bench-Hard
GSM8K	Math	https://github.com/openai/grade-school-math
LongBench	Long-Context	https://github.com/THUDM/LongBench
HumanEval	Programming	https://github.com/openai/human-eval
T-Eval	Agent	https://github.com/open-compass/T-Eval
CHID	Language	https://github.com/chujiezheng/ChID-Dataset
		Holdout Dataset
TyDiQA	Knowledge	https://github.com/google-research-datasets/tydiqa
MATH	Math	https://github.com/hendrycks/math?tab=readme-ov-file

Prevent overfitting and cheating. To mitigate the risk of overfitting on conventional evaluation datasets, we opt to utilize the OpenCompass platform [5] for evaluation purposes. From this platform, we have deliberately curated a diverse set of 7 tasks spanning a wide range of dimensions, along with a set of secret holdout tasks to avoid overfitting. This comprehensive evaluation approach not only guards against overfitting to a specific perspective but also enhances the safety of edge LLMs by uncovering unforeseen challenges and potential pitfalls. To prevent cheating, we ask the participants to submit all their source codes, evaluation log files as well as the model to us. We will then replicate the entire process of the top 3 teams with the highest rank in the first track, to ensure it is repeatable and the final model can be run on a smartphone with 12 GB RAM with the same results.

2.2 Rules and engagement

The rules of this competition aim to ensure fair evaluation and reproducible results. To this end, we require the following:

- Each team is limited to a maximum of five members. Teams must self-certify that no team member is participating in multiple teams for this competition.

- All teams are encouraged to submit their models for a preliminary review. After the review deadline, only three submissions are allowed for each team and the best-performing model will be used to rank the team.
- The use of exclusively C4 and Alpaca datasets is permitted for participation in this competition. The utilization of data or content that violates service agreements or proprietary information of any entity is strictly prohibited. Submissions must refrain from employing any copyrighted or proprietary data, code, or closed-source content.
- All submissions must ensure full reproducibility. The top three teams in each track category, with the highest scoring models, are mandated to provide all essential code and data required for replicating their model. Their submissions must be open-sourced and made publicly accessible after the competition.
- It is not allowed to mix submissions across tracks. Models compressed from pre-trained LLMs cannot be submitted to the training-from-the-scratch track, and vice versa.
- This competition will be run under the honour system. Teams that submit very similar results or copy another team's solution will be disqualified. Violating the spirit of the honour system or taking unfair advantage of the community, even when not against an explicit rule, may result in disqualification and ineligibility for prizes.
- Competition participants can communicate with the organizers to ask questions via various ways such as emails, GitHub issues, and Slack channels. Any updates to rules or deadlines will be timely demonstrated on our website and communicated through our Slack channels.
- Organizers are not allowed to participate in the competition.

2.3 Schedule and readiness

Table 2: Schedule for this competition.

25th June, 2024	Announcement and registration start, "starting kit" will be released
25th July, 2024	Registration deadline and submission open
25th August, 2024	Preliminary review deadline
25th October, 2024	Submission deadline
20th November, 2024	Winners notification
11th December, 2024	In-person workshop

At the time of writing this proposal, the competition website has been created through `https://edge-llms-challenge.github.io/edge-llm-challenge.github.io/`, and part of the relevant information remains to be prepared. The primary part of the "starting kit", i.e., the tool for deploying LLMs on smartphones, is ready, and the rest will be ready by 25th June 2024.

2.4 Competition promotion and incentives

We intend to connect with a more diverse group of people through a variety of channels including social media platforms, academic email lists, professional networks, personalized invitations, Slack servers, and Google groups with special interests in applications of LLMs in various practical disciplines with edge devices. Our potential audience is researchers in various fields such as natural language processing, edge computing and IoT, mobile computing, system and hardware design, privacy and security, and industry practitioners. As NeurIPS encourages diversity, we also target women-in-science audiences to ensure a more equitable scientific data-centric community. We plan to send our announcements, advertisements, and invitations in different language versions and make them more welcoming and inclusive for the diverse audience we are targeting.

We will also offer incentives to encourage individuals to participate and attend our challenges. These incentives include prizes for winning groups in both categories, exclusive offers/discounts or registration-fee scholarships for participating students to attend. For students interested in participating/attending the competition, we will include information of the student volunteer and D&I subsidies program in our announcements and invitations. We will periodically promote our workshop through media advertisements, and we will include engaging visuals to help capture people's attention. The advertisements may include elements like past NeurIPs workshop's success, appealing landmark

scenes of the hosting city, speakers' pictures/biographies, and sponsors. Specifically, for each of the two hardware categories, we will select winning groups for the top three positions, with the following prizes:

- The first-place winning group will receive a $10,000 cash prize,
- The second-place group will receive $5,000, and
- The third-place group will receive $2,000.

Moreover, we will select three student awards per competition category, granted to the highest-ranking groups outside the top 3 positions, that are composed entirely of students. Each student group will receive $1000 travel award to attend the workshop in person. Winners will be announced on the competition website as well as during the in-person workshop at NeurIPS. The two first-place winning teams of each category will be invited to give a 30-minute presentation each during the in-person workshop at NeurIPS 2024, and all members of those teams will be offered a chance to co-author the report-out paper on this competition.

3 Resources

3.1 Resources provided by organizers

The organizers form a diverse and talented team, proficient in edge computing, LLMs, and ML efficiency. Tasks have been allocated efficiently, and all organizers are dedicated to ensuring the seamless operation of the competition from inception to conclusion.

Furthermore, the organizers have enough GPUs and a robust code base to assess submission performance across selected tasks using the OpenCompass platform. We are grateful to our main sponsor for their generous financial backing, which includes funding for monetary awards and the provision of numerous smartphone devices for submission evaluation. We are currently in discussions with additional potential compute sponsors, such as AWS, to secure further support for the competition.

3.2 Support requested

Our competition is self-contained and does not require any special support. However, for the in-person workshop event, we will need a workshop room that can accommodate approximately 400-600 attendees and ten poster boards.

3.3 Organizing team

Shiwei Liu (he/him) 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 main research interest focuses on improving the efficiency and accessibility of LLMs, making them accessible tooling to everyone. 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 Sparsity in Neural Network workshop in ICLR'23. As the lead organizer and competition coordinator, Dr. Liu is responsible for the overall organization and management of the competition. His duties include writing the competition proposal, setting the competition timeline, coordinating communication between co-organizers, ensuring compliance with competition guidelines, and resolving any issues.

- Email: shiwei.liu@maths.ox.ac.uk
- Web page: `https://shiweiliuiiiiiii.github.io/`
- Google Scholar: `https://scholar.google.com/citations?user=73IbXtsAAAAJ&hl=en`

Adriana Fernandez-Lopez (she/her) (PhD 2021, Pompeu Fabra University) is a Postdoctoral Researcher at Meta UK. Her research interests lie in the intersection of computer vision and speech processing, with a focus on using machine intelligence to better understand human behavior. Currently, she is working on compressing and optimizing large foundation models, such as speech recognition models and generative models through the use of sparse networks. Dr. Fernandez-Lopez serves as the platform administrator of our competition. She is responsible for setting up the competition website, maintaining server infrastructure, managing user accounts, and ensuring the smooth operation of the competition platform and the leaderboard.

- Email: afernandezlopez@meta.com
- Web page:`https://adrianafernandez02.github.io/bio.github.io/`
- Google Scholar: `https://scholar.google.com/citations?user=B10dWikAAAAJ&hl=en&oi=ao`

Kai Han (he/him) is currently a researcher with the Huawei Noah's Ark Lab. He received the B.S. and M.S. degrees from Zhejiang University and Peking University respectively, and the Ph.D degree from Institute of Software Chinese Academy of Sciences. His research interests mainly include deep learning, computer vision, and foundation models. He has published over 50 papers in prestigious journals and conferences. He regularly serves as a PC/senior PC member for top conferences, e.g., NeurIPS, ICML, ICLR, CVPR, and ICCV.

- Email: kai.han@huawei.com
- Web page:`https://iamhankai.github.io/`
- Google Scholar: `https://scholar.google.com/citations?user=vThoBVcAAAAJ`

Cong (Callie) Hao (she/her) is an assistant professor in ECE at Georgia Tech, where she currently holds the Sutterfield Family Early Career Professorship. Her primary research interests lie in the joint areas of efficient hardware design, electronic design automation tools, and machine learning algorithms. Dr. Hao has been distinguished with the NSF CAREER award (2024), Intel Rising Star Faculty Award (2023), Amazon Research Award (2022), and Sony Faculty Innovation Award (2022). She also won research awards including the FPL Community Award (2023), Best Paper at IEEE DAC (2023), Best Paper Runner-up at IEEE FCCM (2023), Best Paper Nomination at IEEE ASAP (2022), and Best Paper Award at IEEE GLSVLSI (2021). She has rich experience organizing design competitions. She was the organizer for Design Automation Conference System Design Competition (DAC-SDC) from 2021 to 2023 and was the competition winner from 2018 to 2020. Dr. Hao serves as the baseline method provider, whose responsibility is to develop and provide baseline methods or algorithms that serve as naive solutions for the "starting kit".

- Email: callie.hao@ece.gatech.edu
- Web page: `https://sharclab.ece.gatech.edu/`
- Google Scholar: `https://scholar.google.com/citations?user=fWEIPSUAAAAJ&hl=en`

Rebekka Burkholz (she/her) is a tenure-track faculty member (rank of associate professor) at the Helmholtz Center CISPA, where she leads the relational machine learning group. Their research is funded by the ERC starting grant SPARSE-ML, which aims to democratize deep learning by reducing the associated development and deployment costs of large models. Their main approach is to gain a theoretical understanding of deep learning from a complex network perspective and improve contemporary sparse training algorithms based on these insights. She was awarded the best poster prize at the Computational Cancer Biology (CCB) satellite of RECOMB 2023 and received a lightning talk and best poster prize at Informatics Technology and Cancer Research (ITCR) 2022. Her PhD thesis on systemic risk at ETH Zurich won the Zurich Dissertation Prize and her work on international maize trade received the CSF Best Contribution Award. Dr. Burkholz serves as the Platform Administrator of our competition. She is responsible for setting up the competition website, maintaining server infrastructure, managing user accounts, and ensuring the smooth operation of the competition platform and the leaderboard.

- Email: burkholz@cispa.de

- Web page: https://cispa.de/en/research/groups/burkholz
- Google Scholar: https://scholar.google.ch/citations?user=vkWBb2wAAAAJ&hl

Olga Saukh (she/her) is an associate professor at TU Graz, where she leads the Embedded Learning and Sensing Systems group. She is also a faculty at the Complexity Science Hub. She holds a habilitation degree in Embedded Systems from TU Graz since 2020. She did her postdoctoral training at ETH Zurich in 2010-2016 working in the Computer Engineering and Networks Laboratory. Her PhD thesis received the CONET PhD Academic Award (European Award Competition). Her research focuses on efficient machine learning, engineering and optimization of AI-based systems for resource-constrained devices, covering a range of topics on the intersection of deep learning and embedded / mobile systems. She co-organized multiple events in the past, including the recent IEEE DCoSS 2023 (TPC Co-chair) and IEEE ICPADS 2022 (Track Chair).

- Email: saukh@tugraz.at
- Web page: http://olgasaukh.com/
- Google Scholar: https://scholar.google.com/citations?user=f-MDKlYAAAAJ

Yunhe Wang (he/him) is currently a senior researcher with the Huawei Noah's Ark Lab, where he leads the group of Applied AI Lab. He received the Ph.D. degree from Peking University, China. His research interests mainly include machine learning, computer vision, and efficient deep learning. He has published over 100 papers in prestigious journals and conferences. Many of them are widely applied in industrial products, and received important awards. He regularly serves as a PC/senior PC member for top conferences, e.g., NeurIPS, ICML, ICLR, CVPR, ICCV, IJCAI, and AAAI.

- Email: yunhe.wang@huawei.com
- Web page: https://www.wangyunhe.site/
- Google Scholar: https://scholar.google.com/citations?user=isizOkYAAAAJ

Edoardo M. Ponti (he/him) (PhD 2021, University of Cambridge) is a Lecturer (U.K. Assistant Professor) in Natural Language Processing at the University of Edinburgh and an Affiliated Lecturer at the University of Cambridge. Previously, he was a visiting postdoctoral scholar at Stanford University and a postdoctoral fellow at Mila Montreal and McGill University. His research is centred on modular deep learning, including parameter-efficient fine-tuning, mixtures of experts, and model merging. He received a Google Research Faculty Award and 2 Best Paper Awards at EMNLP 2021 and RepL4NLP 2019. He is currently co-leading an ELIAI grant on Gradient-based Learning of Complex Latent Structures. He is a member of the ELLIS Society and part of the TACL journal editorial team. Dr. Ponti is the baseline method provider for the competition. In this role, he advises on the selection of baselines and evaluation tasks for the competition.

- Email: eponti@ed.ac.uk
- Web page: https://ducdauge.github.io/
- Google Scholar: https://scholar.google.com/citations?user=tklL2q0AAAAJ

Boqian Wu (she/her) is a joint PhD student at the University of Twente, The Netherlands, and the University of Luxembourg, Luxembourg. Her research interests focus on the application of efficient machine learning in healthcare, as well as the exploration of theoretical aspects related to efficient networks. She has publications on sparse neural networks in conferences such as NeurIPS and ICLR, and has served as a PC member for NeurIPS 2023, the SNN workshop in 2022 and 2023. She co-organized a tutorial at IJCAI 2023. Boqian Wu will be one of the evaluators of this competition. She will assess and evaluate the submissions received from participants based on the selected 7 tasks on the OpenCompass platform.

- Email: b.wu@utwente.nl
- Web page: https://people.utwente.nl/b.wu
- Google Scholar: https://scholar.google.com/citations?user=5dACFlcAAAAJ&hl=zh-CN

Ajay Jaiswal (he/him) is a Ph.D. candidate at the Visual Informatics Group, University of Texas at Austin. His research focuses on addressing several fundamental bottlenecks (training, transfer, and inference efficiency, etc.) in the democratization of modern-day neural networks (especially large foundational models and graph neural networks). His research accomplishments are evidenced by several impactful publications in top-tier venues, e.g., NeurIPS, ICML, ICLR, ECCV/ICCV, etc. He is also the winner of the Amazon Science Ph.D. Fellowship 2023. Ajay Jaiswal will be one of the evaluators of this competition. He will assess and evaluate the submissions received from participants based on the selected 7 tasks on the OpenCompass platform.

- Email: ajayjaiswal@utexas.edu
- Web page: `https://ajay1994.github.io/`
- Google Scholar: `https://scholar.google.com/citations?user=I783HxYAAAAJ&hl=en`

Zahra Atashgahi (she/her) is a Ph.D. candidate at the Department of Electrical Engineering, Mathematics and Computer Science (EEMCS), University of Twente, The Netherlands. Her research interests include cost-effective artificial neural networks, sparse neural networks, feature selection, ensemble learning, time series analysis, and healthcare. She has publications on sparse neural networks in journals and conferences, such as Machine Learning, TMLR, ECMLPKDD, ICLR, and NeurIPS. She has served as a PC member of conferences and workshops, including ICML, NeurIPS, AAAI, and the SNN workshop. She co-organized tutorials at ECMLPKDD 2022 and IJCAI 2023, and a workshop at ICLR 2023. Zahra Atashgahi will be one of the evaluators of this competition. She will assess and evaluate the throughput of the submission on a smartphone.

- Email: z.atashgahi@utwente.nl
- Web page: `https://zahraatashgahi.github.io/`
- Google Scholar: `https://scholar.google.com/citations?user=_nmvlmkAAAAJ&hl=en`

Lu Yin (he/him) is a Lecturer (U.K. Assistant Professor) in the Department of Computer Science at the University of Surrey, affiliated with the NICE (Nature Inspired Computing and Engineering) research group. He is also a long-term visiting researcher at Eindhoven University of Technology (TU/e). Previously, he was an Assistant Professor at the University of Aberdeen, served as a Postdoctoral Fellow at TU/e, and was a research scientist/intern at Google's NYC office. His research focuses on AI efficiency, AI for Science, Large Language/Foundation Models. He has published over 20 research papers at leading AI conferences, including ICML, NeurIPS, ICLR, AAAI, and UAI. He maintains long-term and stable research collaborations with renowned industry leaders such as Google, Meta, and Intel.

- Email: l.yin@surrey.ac.uk
- Web page: `https://luuyin.com`
- Google Scholar: `https://scholar.google.com/citations?user=G4Xe1NkAAAAJ`

Tianjin Huang (he/him) is a Lecturer (U.K. Assistant Professor) in the Department of Computer Science at University of Exeter, and a long-term visiting researcher at Eindhoven University of Technology (TU/e). Prior to this, he was a postdoctoral fellow at Eindhoven University of Technology (TU/e), where he also earned his Ph.D. in the Department of Mathematics and Computer Science. His research aims to develop accurate, trustworthy, and efficient deep learning systems, with a primary focus on model efficiency and trustworthiness. Additionally, he is also interested in various AI applications. He has published over 20 research papers within these areas in prestigious conferences and journals such as ICML, NeurIPS, ICLR, LoG, and Information Fusion. He received the Best Paper Award at the inaugural Learning on Graphs (LoG) Conference in 2022.

- Email: t.huang2@exeter.ac.uk
- Web page: `https://tianjinyellow.github.io/`
- Google Scholar: `https://scholar.google.com/citations?user=yFLmPsoAAAAJ&hl=nl`

Andreas Zinonos (he/him) is a PhD candidate researching Artificial Intelligence and Machine Learning at the Department of Computing, Imperial College London, U.K. He has a BSc in Computer Science from University College London, and an MSc in Artificial Intelligence & Machine Learning from Imperial College London. His research interests include deep learning, computer vision, self-supervised learning, speech-processing and generative models. He has publications on audio-visual self-supervised learning in ICASSP, and has presented and tutored a lecture and workshop on self-supervised learning at the AI Tech School 2023 in Warsaw, Poland. It is worth noting that Andreas will be among those evaluating this competition.

- Email: andreas.zinonos18@imperial.ac.uk
- Google Scholar: `https://scholar.google.co.uk/citations?user=JqenOQQAAAAJ&hl=en&oi=ao`

Jared Tanner (he/him) is the Professor of the Mathematics of Information in the Mathematical Institute at the University of Oxford where he leads the Machine Learning and Data Science Research Group. He has previously held faculty positions at the University of Edinburgh and University of Utah as well as postdoctoral positions at Stanford University and University of California at Davis. He received his doctorate at UCLA in Applied Mathematics in 2002. His research focus is on the design, analysis, and application of algorithms for the processing of information. His current focus is understanding how to improve the stability and computational efficiency of deep networks. This includes both theory and experimental work on the design of network activations and wseven structures. He is currently applying these techniques in medical imaging, multispectral sensing, and hardware aware deep learning algorithms. Previous research contributions include theory, algorithms, and applications of compressed sensing, matrix completion, low-rank plus sparse models, and grid free super-resolution.

- Email: tanner@maths.ox.ac.uk
- Web page: `https://people.maths.ox.ac.uk/tanner/`
- Google Scholar: `http://scholar.google.co.uk/citations?user=J7248tkAAAAJ&hl=en`

References

[1] Dataset card for c4 in huggingface. https://huggingface.co/datasets/c4, 2019.

[2] Common crawl website. https://commoncrawl.org, 2023.

[3] Nemo framework user guide. https://docs.nvidia.com/nemo-framework/user-guide/latest/performance.html, 2023.

[4] Phi-2: The surprising power of small language models. https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/, 2023.

[5] Opencompass evaluation. https://opencompass.org.cn/home, 2024.

[6] Smartphone memory: Gen ai upgrades to drive spike in dram demand. https://www.yolegroup.com/technology-outlook/smartphone-memory-gen-ai-upgrades-to-drive-spike-in-dram-demand, 2024.

[7] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, et al. Qwen technical report. arXiv preprint arXiv:2309.16609, 2023.

[8] E. Frantar and D. Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In International Conference on Machine Learning, pages 10323–10337. PMLR, 2023.

[9] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. arXiv preprint arXiv:2210.17323, 2022.

[10] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027, 2020.

[11] Z. Liu, C. Zhao, F. Iandola, C. Lai, Y. Tian, I. Fedorov, Y. Xiong, E. Chang, Y. Shi, R. Krishnamoorthi, et al. Mobilellm: Optimizing sub-billion parameter language models for on-device use cases. arXiv preprint arXiv:2402.14905, 2024.

[12] K. T. Malladi, B. C. Lee, F. A. Nothaft, C. Kozyrakis, K. Periyathambi, and M. Horowitz. Towards energy-proportional datacenter memory with mobile dram. ACM SIGARCH Computer Architecture News, 40(3):37–48, 2012.

[13] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21(140):1–67, 2020.

[14] silk road. alpaca-data-gpt4-chinese. `https://huggingface.co/datasets/silk-road/alpaca-data-gpt4-chinese`, 2023.

[15] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

[16] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

[17] G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, and S. Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In International Conference on Machine Learning, pages 38087–38099. PMLR, 2023.

[18] L. Yin, Y. Wu, Z. Zhang, C.-Y. Hsieh, Y. Wang, Y. Jia, M. Pechenizkiy, Y. Liang, Z. Wang, and S. Liu. Outlier weighed layerwise sparsity (owl): A missing secret sauce for pruning llms to high sparsity. arXiv preprint arXiv:2310.05175, 2023.