Open Science for Foundation Models
ICLR 2025 Workshop Proposal

1 Workshop Summary

Foundation models (FMs) have revolutionized artificial intelligence (AI) research across many domains, enabling rapid adaptation to diverse downstream tasks. These FMs, trained on massive and high-quality datasets, have demonstrated remarkable performance in natural language processing (e.g., BERT [4], GPT [12], Gemini [14]), computer vision (e.g., ViT [5], VQGAN [6]), speech recognition (e.g., Whisper [13]), and multi-modal understanding (e.g., GPT-4o [1], LLaVA [10], QwenVL [2]). Despite these advancements, the scientific transparency and reproducibility of FMs have not kept pace. Proprietary interfaces conceal crucial details, such as training data, architectural design, and development processes, limiting scientific understanding of these models' biases and risks. To bridge this gap, there is a growing need for truly open foundation models that the research community can access and study.

In response, a surge of open science works has emerged to address the issue, encouraging transparency of FMs within the research community. Notable examples include open-access large language models (LLMs) such as Llama [15], Mistral [8], and Qwen [1], as well as extensive pre-training datasets like RedPajama [3] and The Stack [9]. These efforts have democratized access to high-performance models and sparked further innovation. Moreover, several initiatives like OLMo [7] and StarCoder [11] now offer fully transparent models, providing detailed insights into training protocols, intermediate checkpoints, and data processing pipelines. Such transparency is critical to fostering reproducibility and accelerating research across the field.

Therefore, the first **Open Science for Foundation Models (OS-FMs)** workshop aims to bring together a community of researchers committed to open science, reproducible research, and the open-source movement within AI. This workshop seeks contributions that explore key aspects of FMs, such as dataset curation, evaluation methodologies, high-performing models, and efficient implementations. While models have become increasingly large in this era, the workshop promotes the open sharing of both small (e.g., 1B) and large models, as long as their conclusions are based on rigorous scientific experiments. By **emphasizing scientific discovery and open sharing**, the workshop seeks to address the growing inaccessibility of foundation models, ensuring that the benefits of AI advancements are disseminated across the global research community.

Tagline The OS-FMs workshop seeks a discussion on increased transparency in the field by promoting the use of open models and open-source initiatives as an alternative to closed approaches.

2 Submissions

The OS-FMs workshop invites empirical and theoretical research, descriptions, and release of high-quality open datasets, open models, and open-source software implementations, as well as position and survey papers reflecting on how open data, open models, and open-source initiatives can contribute to advances in the field. The possible topics of discussion include (but are not limited to) the following:

- **Open Datasets**: Acquisition, curation, and synthesis of pretraining, instruction, and preference datasets through manual or algorithmic methods. Open access to instruction and preference datasets for alignment research.
- **Open Foundation Models**: Pretraining strategies including data scaling, model architecture, multi-modal, and multi-task pretraining. Learning algorithms such as meta-learning, model fusion, model merging, and continual learning designed for open, scalable models. Inference algorithms like decoding, reasoning, search, and planning, tailored for foundation models.
- **Open Training Protocols**: Training dynamics research on scaling laws, interpretability, complexity analysis, emergent capabilities, and phenomena like grokking. Alignment techniques including prompt tuning, prefix tuning, instruction tuning, and reinforcement learning with human/AI feedback.

[1] https://openai.com/index/hello-gpt-4o/

- **Open Evaluation**: Benchmark development and the creation of transparent evaluation protocols and metrics, including the open sharing of benchmark datasets and evaluation results across different foundation models.
- **Open Compute Efficiency Techniques**: Focus on model distillation, compression, quantization, and optimizing attention or memory mechanisms for improved compute efficiency in open foundation models.
- **Open Multi-Modal Foundation Models**: Expanding to modalities like vision, audio, and multi-modal foundation models, with extra emphasis on underexplored areas such as chemistry, medicine, and education.
- **Open Interactive and Agent Systems**: Open development of conversational AI, interactive learning models, multi-agent systems, and integration with external tools and APIs.
- **Open Replication of Proprietary Systems**: Efforts to replicate and openly share foundation models and systems that were previously proprietary, ensuring transparency and reproducibility for broader research and development.

Submission formats

The workshop will accept research papers and industrial papers, from preliminary research results and visionary papers to full-length papers. Submissions should follow the ICLR proceedings format and choose the suitable categories as follows:

- Short paper: up to 6 pages + references and appendix.
- Regular paper: up to 9 pages + references and appendix.

All accepted papers will be presented as posters. We will select around 3 papers for short oral presentations and 2 papers for outstanding paper awards with potential cash incentives.

Novelty and Openness

For novelty, the workshop does not accept submissions that have previously been published at ICLR or other machine learning or related venues. For openness, we encourage submissions with sufficient open-source resources (e.g., checkpoint, code, data, training details). Accepted papers will be published on the website but the workshop is non-archival.

Review process

We will use OpenReview to organize the review cycle. Each paper will be reviewed by three PC members. Besides, we will provide review guidelines to express our expectations on the reviews. Authors will mark conflicts of interest with the workshop organizers and the program committee, and reviewing will be handled accordingly. Organizers are allowed to submit their own works to the workshop as OpenReview disables the visibility of PC assignments and reviews, while another organizer can handle the decision-making for that submission.

Submission timeline

- Submission open: 22 January 2025
- Submission deadline: 3 February 2025
- Notifications of accept/reject: 5 March 2025
- Camera-ready, slides, and recording upload: 5 April 2025

3 Workshop Schedule and Organization Details

Schedule

The program will last for one full day (the host day could be accommodated to the conference schedule), and the proposed schedule details are suggested as follows:

Program Elements

Estimated interest The openness of foundation model research generally perceives a rising interest in the community, in both academia and industry (large companies, public organizations, and startups). According to the observation of the organization committee, the highly active discussion on the degree of open-source and reproducibility of the latest published foundation models research in the communities suggests the demand of the promotion of more openness. This trend also reflects at the popularity of the recent open science theme track at ACL 2024 (38/55 acceptance/submission numbers). To sum up, we estimate the number of participants in the OS-FMs workshop as 100-150 according to our past experience.

Table 1: Workshop Schedule

Time	Arrangement
08:00–09:00	**Registration**
09:00–09:10	**Opening Remarks**: Welcome address and workshop overview
09:10–09:40	**Keynote Talk 1**: Stella Biderman, EleutherAI (confirmed)
09:40–10:10	**Keynote Talk 2**: Loubna Ben Allal, Hugging Face (confirmed)
10:10–10:40	**Oral Presentations** from submissions
10:40–11:10	*Coffee Break*
11:10–11:40	**Keynote Talk 3**: Zhengzhong Liu, Petuum (confirmed)
11:40–12:30	**Poster Session 1**
12:30–13:30	*Lunch Break*
13:30–14:00	**Keynote Talk 4**: Wenhao Huang, ByteDance (confirmed)
14:00–14:30	**Keynote Talk 5**: Shayne Longpre, MIT (confirmed)
14:30–15:00	**Keynote Talk 6**: Luca Soldaini, Allen Institute for AI (confirmed)
15:00–16:00	**Panel Discussion**: How could openness benefit the research of foundation models and how far are we from fully open science?
16:00–16:30	*Coffee Break*
16:30–17:20	**Poster Session 2**
17:20–17:30	**Award Announcement** and **Closing Remarks**

Oral Presentations and Poster Sessions To inspire a comprehensive discussion on the latest effort on the open foundation models, we plan to arrange three 10-minute oral presentations selected from the accepted submissions, and two poster demonstration sessions for all the accepted papers.

Invited Keynote Talks With the invited talks (30 minutes each), we aim to invite the researchers contributed to the open science of foundation models from various aspects, demonstrating the necessity and challenges of promoting the open research of foundation models in depth. We are proud of the diversity of keynote speakers regarding expertise, demographics, and affiliations, which could provide the audience with first-hand experiences rooted in a wide spectrum of backgrounds. The organizers monitor the time of speakers and ensure that there is sufficient time for discussion. All invited speakers have confirmed to give their talks in person. We provide the details of the invited keynote speakers at §3.

Details of Panel Discussion Following the main sessions (keynote talks, contributed talks, and posters), we will host a panel discussion session, where the invited speakers will discuss the centering topics of the workshop. First, we will review the role of open science in the development history of modern foundation models. Following that, the speakers will further discuss and provide their opinions on the existing dilemma and potential solutions for advancing the openness of future FM research, which could be naturally derived from true experiences thanks to their diverse backgrounds from academia to industry.

Accessibility The workshop will be held in a **hybrid** format, as we understand the potentially unexpected constraints of the contributors may have (personal, health, financial, etc.). This hybrid format combines on-site participation with live streaming, ensuring broader accessibility. Besides, we will require authors who cannot attend in person to upload a poster to the workshop website and will encourage them to optionally provide a brief 3-minute video presentation. Moreover, to maximize engagement, we have implemented a comprehensive communication strategy, which encompasses pre-event promotion, real-time interaction during the workshop, and post-event follow-up. Additionally, key research concepts will be shared via our workshop website and popular social media channels to attract a wide-ranging audience.

Details of Keynote Speakers

Stella Biderman (confirmed) is the head of research at EleutherAI, an open-source AI research lab that has revolutionized open access to large language models. Her academic journey began with a Bachelor's degree from the University of Chicago, followed by a Master's degree from the Georgia Institute of Technology. She is best known for her work on democratizing LLMs including open-source models and datasets, (e.g., BLOOM, Pythia, GPT-NeoX-20B, The Pile). Her work on publicly available datasets and evaluation frameworks has become an integral part of training foundation models.

Loubna Ben Allal (confirmed) is a Machine Learning Engineer at Hugging Face, where she spearheads initiatives on training small Language Models (SmolLM) and developing pre-training datasets such as Cosmopedia and FineWeb-Edu. Her expertise extends to large-scale language models, evidenced by her previous role as a core team member of BigCode, where she contributed significantly to The Stack dataset—the largest open dataset of source code—and the development of the StarCoder and StarCoder2 model families. Loubna holds a Master's degree in MVA from ENS Paris Saclay and an engineering degree with a major in Mathematics from École des Mines de Nancy. Originally from Midelt, Morocco, Loubna now applies her diverse background and technical skills to advance the field of natural language processing and machine learning in Paris, France.

Zhengzhong Liu (confirmed) , also known as Hector, is the Head of Engineering at Petuum and a leading figure in the field of Natural Language Processing (NLP) and Large Language Models (LLMs). He earned his PhD from Carnegie Mellon University, where he worked under the guidance of Professors Teruko Mitamura and Eduard Hovy. Dr. Liu's research spans a wide range of NLP topics, from event semantics and anaphora resolution to the cutting-edge development of LLMs. He is at the forefront of the LLM360 open-source initiative, which aims to make LLM training processes more transparent and accessible. His work on models like Amber and CrystalCoder showcases his commitment to advancing open-source AI. Dr. Liu's expertise extends to integrating linguistic insights with machine learning techniques, as evidenced by his contributions to semantic web problems and information retrieval. His current focus on understanding how LLMs learn knowledge and abilities through next-token prediction tasks places him at the cutting edge of AI research. Dr. Liu's work consistently demonstrates a balance between theoretical depth and practical application, making him a valuable contributor to both academic and industrial AI advancements.

Luca Soldaini (confirmed) is a senior research scientist at the Allen Institute for AI, where he contributes to the Semantic Scholar and OLMo teams. His research focuses on best practices for the curation and exploration of large corpora, particularly in the context of Large Language Models (LLMs), and efficient domain adaptation in Information Retrieval (IR). Soldaini co-leads the data curation team for OLMo, AI2's state-of-the-art language model, and has been instrumental in developing Dolma, an open dataset of 3 trillion tokens for language model pretraining. His recent work spans diverse areas including cross-language information retrieval, the intersection of LLMs and IR systems, and multimodal PDF processing. He is also an organizer at Queer In AI, where he advocates for participatory AI and community-led initiatives. His contributions have been recognized with best paper and demo awards at prestigious conferences such as FAccT and EMNLP. Prior to joining AI2, He was a senior applied scientist at Amazon Alexa and completed his Ph.D. in computer science at Georgetown University in 2018.

Wenhao Huang (confirmed) is a researcher at ByteDance. He was a co-founder of 01.AI, responsible for pre-training and multimodal models. He holds a Ph.D. from Peking University and has served as the technical lead of the Health Computing Research Center and Innovation Application Laboratory at the Beijing Academy of Artificial Intelligence, as well as a lead researcher at Microsoft Research Asia. In recent years, he mainly focused on research on large language models and multimodal models, overseeing the training and release of 01.AI's large model Yi-large, as well as open-source models such as Yi-34B, achieving excellent results on platforms like LMSYS and Hugging Face's LLM leaderboard. Prior to this, he had extensive experience in empowering enterprise intelligence and financial investment using artificial intelligence technologies. His research outcomes have been applied to products such as Microsoft's natural language understanding platform LUIS, Office, Teams, and Bot Framework, with a user base exceeding 3 billion. He has authored over fifty papers in the field of artificial intelligence, including venues like ICLR, AAAI, and CVPR. He recently gave a talk on "Methodology of large model training and practice of Yi-Large" at the BAAI conference 2024.

Shayne Longpre (confirmed) is a PhD Candidate at MIT, focusing on the intersection of AI and policy, with particular emphasis on the responsible training, evaluation, and governance of general-purpose AI systems. His research has garnered significant recognition, including Best Paper Awards from ACL 2024 and NAACL 2024, and has been featured in prominent media outlets such as the New York Times, Washington Post, and MIT Tech Review. Longpre leads the Data Provenance Initiative and spearheaded the Open Letter on A Safe Harbor for Independent AI Evaluation & Red Teaming. His contributions to the field include work on major language models like Bloom, Aya, and Flan-T5/PaLM. Prior to his doctoral studies, Longpre was a Google Brain Student Researcher and held research positions at Stanford NLP lab and Salesforce Research. He holds a BA in Economics and an MS in Computer Science from Stanford University. Longpre's current work addresses critical issues in AI ethics, transparency, and societal impact, making him a leading voice in shaping the future of AI governance and development.

4 Diversity and Inclusion

Topics: Our workshop embraces a broad spectrum of foundation models including vision, language, audio, etc. The call for papers encompasses cutting-edge research including pretraining, alignment, evaluation, data, etc, which reflects our commitment to comprehensive coverage. We encourage contributions from various academic and cultural backgrounds to foster rich discussions.

Speakers: Our invited speakers come from various institutions, including EleutherAI, MIT, Allen AI, HuggingFace, Petuum, and ByteDance. Their roles span academia and industry, with notable contributions to open science of foundation models like OLMo, LLM360, Yi-Large, FineWeb-Edu, Bloom, The Pile, etc. Additionally, their research areas are also diverse, including open science on pretraining, alignment, data, evaluation, and so on.

Organization Committee: The committee comprises researchers from diverse demographics and experiences, balancing gender, seniority, affiliations, and research areas. Among the eight organizers, two identify as female and six as male, with roles ranging from academic institutions (e.g., Stanford University, University of Manchester, University of Southern California, University of California, Santa Barbara) and industry companies (e.g., Sea AI Lab, ByteDance). All the organizers have worked as organizers or speakers in AI-related workshops.

Attendees: We strive to create a welcoming environment for attendees from diverse backgrounds, facilitating discussions that bridge various perspectives on Foundation Models. To promote broader participation, we will offer a limited number of travel grants to support attendees from underrepresented groups, including but not limited to individuals with disabilities, those from developing countries, and early-career researchers facing financial constraints.

5 Organization Committee

Organizers

Many of the organizers have previous experience with workshops and seminars organization, including experience in organizing the Table Representation Learning Workshop series at NeurIPS, the Machine Learning for Data Workshop at ICML, and Online Workshop on Music & Audio Processing at HKUST. All the organizers will be able to attend the workshop for organizing the workshop and hosting panel discussions in person except for unexpected circumstances. Based on our previous experience, we estimate this in-person workshop in ICLR will have 100-150 attendees. Our previous collective organizing experience can provide substantial evidence for hosting the proposed first OS-FMs workshop.

Jiaheng Liu (buaaljiaheng@gmail.com) is a research scientist at Alibaba Group and he will join Nanjing University as an assistant professor in December 2024. He received his PhD in Computer Science from Beihang University in 2023, advised by Professors Ke Xu and Dong Xu. His current research mainly focuses on Large Language Models (LLMs), with contributions in pre-training, alignment, and open science of LLMs. He has published 40+ top conference/journal papers (ICLR, NeurIPS, ICML, ACL, CVPR, etc.). Recently, he has served as the core author of several well-known projects, including the fully transparent bilingual LLM (MAP-Neo-7B), the high-performance transparent code LLM (OpenCoder), the domain-specific scaling law (D-CPT Law), the Emulated Disalignment method, etc, where the Emulated Disalignment obtains the ACL 2024 Outstanding Paper Award. Besides, he is also one of the co-founders of the M-A-P community, a popular open-source community in China. Additionally, he has served as the invited speaker of the 1st International Workshop on Generalizing from Limited Resources in the Open World at IJCAI 2023.

Riza Batista-Navarro (riza.batista@manchester.ac.uk) is a senior lecturer (equivalent to associate professor) at the School of Computer Science, University of Manchester. She holds a PhD in computer science from the University of Manchester, specializing in biomedical text mining. Batista-Navarro also earned her MSc and BSc in Computer Science from the University of the Philippines Diliman. Her research interests focus on text mining, natural language processing, and machine learning. Batista-Navarro's work has garnered significant recognition in the field, including the prestigious Best Paper Award at ACL2024. Batista-Navarro recently gave a talk at Digital Futures, about applying her expertise to raise awareness about the carbon footprint of home cooking, demonstrating the practical applications of computer science in addressing real-world environmental challenges.

Qian Liu (liuqian.sea@gmail.com, **Local Organizer**) is a Research Scientist at Sea AI Lab based in Singapore. Before joining Sea AI Lab, he was a joint Ph.D. candidate at Beihang University and Microsoft Research Asia. His research interests encompass tabular data, code generation, and natural language reasoning. He has published several papers in top conferences, with notable works including TAPEX, POET, and StarCoder. Qian Liu has received several awards such as the KAUST AI Rising Star in 2024, and was nominated for the Baidu Scholarship in 2020. Additionally, he was one of the co-founders of the MLNLP community, a popular NLP community in China. Qian has served as the organizer for the 3rd Table Representation Learning workshop at NeurIPS 2024 and is also a co-organizer for the Deep Learning for Code workshop under review at ICLR.

Niklas Muennighoff (n.muennighoff@gmail.com) is a first-year PhD student at Stanford University, working under the guidance of Professors Percy Liang and Tatsunori Hashimoto. His research focuses on large language models (LLMs), with particular emphasis on improving their capabilities and usefulness. Niklas's work spans several critical areas in AI, including LLM pretraining, alignment, instruction tuning, and embedding/retrieval systems. His contributions have been recognized with multiple prestigious awards, including the NeurIPS 2023 Outstanding Paper Runner-Up Award and the ACL 2024 Best Theme Paper Award. Prior to Stanford, Niklas completed his Bachelor's degree at Peking University. His collaborative approach to research has led to partnerships with Contextual AI and AI2, and his work has significantly impacted the field, as evidenced by his co-authorship of one of the most influential ACL papers. Niklas's commitment to advancing AI technology is matched by his enthusiasm for mentoring and collaborating with fellow researchers.

Ge Zhang (zhangge.eli@bytedance.com) is a co-founder of M-A-P community and was one of the entrepreneur team of 01.AI, where he leads the open-source of MAP-Neo Series, CT-LLM, and participates into the development of Yi-Series Model. His research focuses on large-scale pretrain of the Foundation Model, LLM evaluation, and AI Ethics. He leads the data curation of MATRIX, the state-of-the-art bilingual transparent pretrain corpus, co-leads the pretrain of a series of state-of-the-art music pretrain models, including Music2Vec, MERT, and MUPT, and participates in development of several well-known multimodal large language models, including AnyGPT, Yi-VL, and MIO, and several domain-specific LLM, including MAmmoTH, OpenCodeInterpreter, and ChatMusician. He also serves as a core member of developing several key metrics of LLM, including MMMU, MMMU-pro, CMMMU, and MMLU-pro.

Yizhi Li (yizhi.li-2@manchester.ac.uk) is a PhD candidate at the University of Manchester, working under the supervision of Prof. Chenghua Lin. As a co-founder of the Multimodal Art Projection (M-A-P) research community, Yizhi has made contributions to the fields of music modeling, fairness in NLP, and vision-language modeling. His work has been recognized at conferences, with multiple papers accepted at ACL, ICLR, NeurIPS, EMNLP, and ISMIR. Yizhi's research has garnered attention both in academic circles and beyond, with his open-source M-A-P models achieving over 50K monthly downloads on Hugging Face. Yizhi's commitment to the field extends to academic service, where he serves as a reviewer for major conferences in computational linguistics and music information retrieval. He recently presented his work on multimodal foundation models and received the best student pitch award at the MultimodalAI'23 workshop.

Xinyi Wang (xinyi_wang@ucsb.edu) is a final-year Ph.D. candidate in the Computer Science department at the University of California, Santa Barbara, advised by Prof. William Yang Wang. Her research focuses on developing a principled understanding of deep learning models, especially large language models, from a Bayesian or causal perspective, with the goal of improving their capabilities, addressing their limitations, and optimizing their application across diverse domains. She has published 7 first-authored papers at leading ML/AI conferences including NeurIPS, ICLR, ICML, COLM, and AISTATS, and more than 10 other co-authored papers, with more than 1000 citations on Google Scholar. She is also a recipient of the J.P. Morgan AI PhD Fellowship. She is an alumnus of the Hong Kong University of Science and Technology, and she has also interned at Microsoft Research Montreal and MIT-IBM Watson Lab before.

Willie Neiswanger (neiswang@usc.edu) is an Assistant Professor of Computer Science at the University of Southern California (USC), in the Viterbi School and School of Advanced Computing. He works on open-source large language models, generative models, and their applications in science and engineering applications. He has co-organized workshops at NeurIPS (NeurIPS 2019, 2023), and ICML (ICML 2020, 2021, 2022). Previously, he completed his Ph.D. at Carnegie Mellon University, and postdoc and Stanford University.

Program Committee

This workshop provides an excellent opportunity to get feedback from different perspectives. We aim to have every submission reviewed by Program Committee members from at least two different research areas. We will have a Program Committee with experts from within and outside our networks, among which the following have already accepted our invitations. The Program Committee is as follows:

Chenyang Zhao (UC Los Angeles), Wenda Li (University of Michigan), Yuanxing Zhang (Kwai Tech), Zelei Chen (Northwestern University), Yue Zhang (ByteDance), Hanqing Wang (ShanghaiTech University), Chujie Zhang (Tsinghua University), Haoran Wang (Tsinghua University), Zekun Wang (ByteDance), Chenchen Zhang (Tencent), Jian Yang (Alibaba, Qwen), Wangchunshu Zhou (OPPO), Zili Wang (Inf Tech), Xinrun Du (01.AI), Tianyu Zheng (01.AI), Jie Liu (CUHK), Qian Liu (Sea AI Lab), Yizhi Li (University of Manchester), Xinyi Wang (UC Santa Barbara), Ge Zhang (ByteDance), Tara Kaul (University of Manchester), Yiming Liang (UCAS), Wei Fan (University of Oxford), Yanjun Shao (Yale University), Yanan Ma (University of Manchester), Xingwei Qu (University of Manchester), Hangyu Guo (Alibaba), Siwei Wu (University of Manchester), Shuyue Guo (Alibaba), Haoran Que (Shanghai AI Lab), Michael Saxon (UC Santa Barbara), Alon Albalak (SynthLabs), Antonis Antoniades (UC Santa Barbara), Alfonso Amayuelas (UC Santa Barbara), He Zhu (Alibaba), Meng Cao (MBZUAI), Yinghui Li (Tsinghua University), Tom Palczewski (SAP), Wen-Ding Li (Cornell University), Xinyuan Lu (NUS), Tan Yu (NVIDIA), Hanhua Hong (University of Manchester)

References

[1] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, and T. Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[2] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[3] T. Computer. Redpajama: an open dataset for training large language models, 2023.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[6] P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis, 2020.

[7] D. Groeneveld, I. Beltagy, E. Walsh, A. Bhagia, R. Kinney, O. Tafjord, A. Jha, H. Ivison, I. Magnusson, Y. Wang, S. Arora, D. Atkinson, R. Authur, K. Chandu, A. Cohan, J. Dumas, Y. Elazar, Y. Gu, J. Hessel, T. Khot, W. Merrill, J. Morrison, N. Muennighoff, A. Naik, C. Nam, M. Peters, V. Pyatkin, A. Ravichander, D. Schwenk, S. Shah, W. Smith, E. Strubell, N. Subramani, M. Wortsman, P. Dasigi, N. Lambert, K. Richardson, L. Zettlemoyer, J. Dodge, K. Lo, L. Soldaini, N. Smith, and H. Hajishirzi. OLMo: Accelerating the science of language models. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics.

[8] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b. *ArXiv*, abs/2310.06825, 2023.

[9] D. Kocetkov, R. Li, L. B. Allal, J. Li, C. Mou, Y. Jernite, M. Mitchell, C. M. Ferrandis, S. Hughes, T. Wolf, D. Bahdanau, L. von Werra, and H. de Vries. The stack: 3 TB of permissively licensed source code. *Trans. Mach. Learn. Res.*, 2023, 2023.

[10] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning, 2023.

[11] A. Lozhkov, R. Li, L. B. Allal, F. Cassano, J. Lamy-Poirier, N. Tazi, A. Tang, D. Pykhtar, J. Liu, Y. Wei, et al. Starcoder 2 and the stack v2: The next generation. *arXiv preprint arXiv:2402.19173*, 2024.

[12] OpenAI. Gpt-4 technical report. *PREPRINT*, 2023.

[13] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR, 23–29 Jul 2023.

[14] G. Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv: 2312.11805*, 2023.

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