Modular, Collaborative and Decentralized Deep Learning

1 Towards Modular, Collaborative and Decentralized Machine Learning

While the success of large-scale deep learning models has hinged on the *"bigger is better"* approach – scaling model size and training data [Sutton, 2019, Kaplan et al., 2020, Henighan et al., 2020, Hoffmann et al., 2022] – this paradigm may rapidly be reaching an inflection point. Beyond the prohibitive cost of training and maintaining gigantic models, this approach exposes and exacerbates inherent flaws in the current design philosophy of machine learning systems.

One of the most glaring contradictions lies in the development life cycle of these models which, once deprecated, are simply discarded in favor of new ones and are generally trained from scratch. This unsustainable practice stems from the fact that models are currently built and trained as *generalist black-box monolithic systems* where functionalities and emerging capabilities are intertwined in their parameters and any attempt to change a specific aspect can have unpredictable and potentially disastrous consequences for the entire model's performance (e.g., *catastrophic forgetting* [McCloskey and Cohen, 1989, French, 1999]).

In stark contrast, a fundamental principle in software development is the organization of code into **modular components** [Parnas, 1972]. This allows developers to import modules and seamlessly integrate new functionalities, improving code reusability and maintainability. Similarly, biological systems provide compelling evidence for the benefits of modularity and functional specialization [Fodor, 1983, Ballard, 1986], such as rapid adaptation to new environments and resilience to perturbations [Wagner et al., 2005]. Despite these clear benefits, modular approaches are rarely applied in the development of machine learning models [Pfeiffer et al., 2023], presenting significant opportunities for innovation.

> This workshop envisions a future where deep learning models are built with **modular design** and **functional specialization** in mind, where sub-parts of the model could be trained independently, unlocking two key capabilities:
>
> - **Asynchronous training and incremental updates**: functional specialization enables efficient, continuous improvement of individual modules, reducing interference, enhancing maintainability and interpretability, and enabling collaborations.
> - **Composability for cross-task generalization**: The modular design enables *post hoc* composition of different modules to adapt to new tasks and domains, promoting reusability and systematic generalization.

Scope and Topics: The scope of this workshop covers all methods enabling collaborative development of modular models. This includes mixture-of-experts where each expert can be independently trained, decentralized training to share regularly information between experts, and upcycling to re-use existing models. In particular, relevant research for this workshop include:

- Mixture-of-Experts [Eigen et al., 2014, Shazeer et al., 2017],
- Routing of Experts (MoErging) [Zhou et al., 2022, Douillard et al., 2024, Filippova et al., 2024, Yadav et al., 2024a];

Preprint. Under review.

- Model Merging [Wortsman et al., 2022, Ramé et al., 2023, Yadav et al., 2024b,c, Kandpal et al., 2023];
- Decentralized Training [McMahan et al., 2023, Douillard et al., 2024, Jaghouar et al., 2024, Diskin et al., 2021];
- Upcycling [Komatsuzaki et al., 2023, Sukhbaatar et al., 2024],
- Adaptive Architectures [Devvrit et al., 2023, Raposo et al., 2024].

Objective of the Workshop: This workshop has two aims. The first is to raise awareness of the wide variety of modular networks, moving beyond established paradigms such as Sparse Mixture-of-Experts [Shazeer et al., 2017]. The second is to bring together more recent and different aspects of modularity (MoE, model merging, parameter-efficient modules, decentralized training, *etc.*) to foster new intersectional ideas. Ultimately, this workshop seeks to contribute to a future where model development is decentralized and collaborative [Kandpal et al., 2023, Raffel, 2023]. The feasibility of such a future is supported by the widespread sharing of models on HuggingFace's Model hub[1] among other platforms. We want to push further and emphasize how researchers can collaborate by combining and extending existing models (*e.g.* model merging and upcycling) and training sub-parts of a model (*e.g.* decentralized modular training) independently.

Impact: The high cost of training large-scale models has kept their development out of the hands of most researchers. Modularity has the potential to democratize the development of machine learning models, shifting from the current centralized paradigm towards a collaborative ecosystem. By enabling independent training and the combination of specialized modules, research institutions and smaller teams could build powerful models by sharing their knowledge and resources. This would also pave the way for more flexible and efficient models, with modularity facilitating updates, customization, and reduced training costs. Finally, the ability to analyze and replace specific modules would dramatically improve model interpretability and trustworthiness, fostering rigorous evaluation and accountability.

2 Planned Activities and Timing

2.1 Submission Timeline

The workshop invites submissions of novel research in two formats: full papers (maximum eight pages) and short papers (maximum four pages). A call for papers will be issued in December with a deadline in early February. Submissions must adhere to the provided ICLR 2025 LaTeX style files[2]. Only original work not previously at any archival venue (including conferences and journals, such as ICLR 2025) will be considered. Preprints on arXiv are permissible; details will be clarified in the call for papers. The workshop is non-archival meaning that these paper or their extended versions can be submitted to conferences and journals in the future. A double-blind review process, employing the OpenReview platform with multiple reviewers and a meta-reviewer, will be implemented to manage potential conflicts of interest. To allow a two-week preparation period following ICLR 2025 decisions and a four-week review period, acceptance notifications will be issued by March 5th.

- ICLR decision notification: **January 22th, 2025**
- Workshop submission deadline: **February 3rd, 2025**
- Workshop Acceptance notification: **March 5th, 2025**

2.2 Virtual Access to Workshop Materials and Outcome

Recognizing the challenges posed by varying time zones in a hybrid meeting format, we will incorporate a blend of synchronous and asynchronous activities to ensure wide participation. Specifically, we will ask authors of accepted papers to provide short pre-recorded videos in advance, enabling registered attendees to access the content flexibly. Live sessions, such as debate discussions and Q&A segments for invited talks or spotlights, will be facilitated through platforms like sli.do. We will also livestream all talks on YouTube.

[1] huggingface.co/models

[2] https://github.com/ICLR/Master-Template/raw/master/iclr2025.zip

2.3 Estimated Attendance

We expect more than 100 submissions and we will accept around 40 papers. More than 150 people are expected to attend in-person and more than 100 virtually. We plan to leverage platforms like Twitter along with other media outlets to publicize our workshop. These platforms will be instrumental in disseminating information about the conference, as well as fostering an online academic dialogue (e.g. spontaneous paper reading groups initiated by the community). We plan to use the diverse networks of the workshop organizers to find relevant reviewers for the conference. We will also explore the creation of a Slack or Discord community to bring together different participants interested in this topic.

2.4 Opportunity for Discussion

The workshop will provide an opportunity for discussion through two main venues. First, the standard poster sessions will allow for discussion among all workshop attendees in the usual manner. Second, we'd like to give the opportunity for our speakers to have a debate. Specifically, instead of semi-directed panel as is usually common in workshops, we plan instead to draw a list of binary questions, all related to modularity, and ask speakers to debate one side vs the other. By doing so, we hope to create more animated dialogs and encourage everyone to reconsider their positions.

2.5 Tentative Schedule

This workshop will feature a novel format for invited talks, pairing junior researchers (PhD students and recent graduates) with senior colleagues working in related areas. Junior speakers will give an "opener" talk that presents detailed technical aspects of their research (15-minute presentations), complemented by 30-minute presentations from senior speakers providing broader context, integrated overviews, and future research directions. While this pairing will be prioritized, topic alignment may necessitate deviations from this structure in certain instances. The program will include four 15-minute short invited talks, five 30-minute long invited talks, and a 50-minute debate. Furthermore, three 10-minute spotlight talks will showcase selected submissions. Two 45-minute poster sessions will provide opportunities for informal interaction and scholarly exchange.

3 Invited Speakers

We already confirmed several speakers listed below. We have paired junior and senior speakers wherever applicable based on the topic overlap. We keep some open slots to invite more speakers later during the year from different geographic locations.

1. Decentralized Modularity
   - [15min; Confirmed] Olga Golovneva (Research Scientist at Meta)
   - [30min; Confirmed] Marc'Aurelio Ranzato (Research Scientist at Google DeepMind)
2. Decentralized training across the world
   - [30min; Confirmed] Sami Jaghouar (Founding Engineer at PrimeIntellect) & Max Ryabinin (Distinguished Scientist at TogetherAI)
3. Model MoErging and Merging
   - [15min; Confirmed] Jonas Pfeiffer (Research Scientist at Google DeepMind)
   - [30min; Confirmed] Alessandro Sordoni (Research Scientist at Microsoft Research)
4. Adaptive and Elastic Models
   - [30min; Confirmed] Sneha Kudugunta (PhD student at University of Washington)

4 Diversity commitment

Our commitment to diversity extends across various dimensions, including but not limited to ethnicity, age, gender, background, scientific expertise, and knowledge.

Time	Event	Duration
08:50 – 09:00	Opening Remarks	10 min
09:00 – 09:15	Short Invited Talk #1	15 min
09:15 – 09:45	Long Invited Talk #1	30 min
09:45 – 10:00	Short Invited Talk #2	15 min
10:00 – 10:30	Long Invited Talk #2	30 min
10:30 – 10:45	Coffee Break	15 min
10:45 – 10:55	Contributed Talk #1	10 min
10:55 – 11:05	Contributed Talk #2	10 min
11:05 – 11:15	Contributed Talk #3	10 min
11:15 – 12:30	Poster Session #1	45 min
12:30 – 14:00	Lunch Break	90 min
14:00 – 14:15	Short Invited Talk #3	15 min
14:15 – 14:45	Long Invited Talk #3	30 min
14:45 – 15:00	Short Invited Talk #4	15 min
15:00 – 15:30	Long Invited Talk #4	30 min
15:30 – 16:00	Long Invited Talk #5	30 min
16:00 – 16:15	Coffee Break	15 min
16:15 – 17:00	Poster Session #2	45 min
17:00 – 17:50	Debates	50 min
17:50 – 18:00	Closing Remarks	10 min

Table 1: Tentative Workshop Schedule

Gender Diversity: Our effort to ensure balanced gender representation can be seen in the list of speakers and organizers. We have ensured that both male and female organizers are actively involved, contributing their unique perspectives to the discussions. And our speakers are also balanced in terms of gender: Among 8 speakers, 3 are female and 5 are male.

Geographic and Institutional Representation: We involve speakers and organizers from diverse geographic locations and represent a mix of universities and research institutions. Organizers are affiliated with different institutions, including the University of North Carolina at Chapel Hill, University of Toronto, University of Cambridge, and Google DeepMind. Speakers also cover both academic institutions and industry companies and startups in different countries. We are targeting geographic and institutional diversity, which ensures that multiple educational cultures and methodologies are represented.

Range of Scientific Seniority: We bring together a range of scientific seniorities. This ensures that attendees benefit from both fresh perspectives and seasoned insights. Among organizers, there are faculty members, postdoctoral fellows, research scientists, and PhD students. Invited speakers also range from PhD students embarking on their academic journeys to research scientists and established names in the field, offering different experiences, viewpoints, perspectives, and beyond.

5 Previous Related Workshops

This is the first iteration of the Modular, Collaborative, and Decentralized Deep Learning Workshop. Many workshops and tutorials on related themes have been held in recent years, including:

- **Compositional Learning**: Compositional Learning: Perspectives, Methods, and Paths Forward (NeurIPS 2024) This workshop is closely related to our workshop, as both aim for compositional generalization and involves using modular components. However, our workshop also emphasizes large-scale training and decentralized collaboration, which may lead to new model development paradigms.

- **Decentralized Learning and Federated Learning**: Workshop on Decentralized and Collaborative Learning (MLSys 2023), Cross-Community Federated Learning: Algorithms, Systems and Co-designs (MLSys 2022), International Workshop on Federated Foundation Models (NeurIPS 2024), International Workshop on Federated Learning in the Age of Foundation Models (NeurIPS 2023 Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities (ICML 2023) Flower AI Summit (2024) These workshops share our goal of building models in a collaborative and decentralized way. However, we focus on enabling asynchronous modular updates, giving the participants more autonomy in how to train their modules, and upcycling existing open models. Federated Learning, on the other hand, expects the participants to adhere to a consistent learning algorithm. Due to the need to handle cross-silo and cross-device settings, Federated Learning generally requires more communication effort than the scope of our workshop.
- **Continual Learning and Transfer Learning**: Scalable Continual Learning for Lifelong Foundation Models (NeurIPS 2024) Workshop on Continual Learning in Computer Vision (CVPR 2023) Transfer Learning for NLP Workshop (NeurIPS 2022) Both Continual and Transfer Learning have deep connection to our workshop. Many challenges and approaches are shared. However, they are structured top-down, iterating solutions to the well-defined Continual Learning and Transfer Learning problems. In contrast, our scope definition is bottom-up, stemming from the existing development practices of training, sharing, tweaking, and combining modules. Our workshop seeks to systemize and coordinate haphazard and small-scale development activities.
- **Mixture of Experts**: Mixture-of-Experts in the Era of LLMs: A New Odyssey (ICML 2024) Mixture of Experts architectures share similarities with some classes of modular models. Our workshop places Mixture of Experts in the context of other modularity approaches, such as merging, routing, and adaptive architecture, creating a more comprehensive picture of modularity. In addition, Mixture of Experts is typically trained in a centralized setting, and its modularity is in support of computation. In contrast, we leverage modularity to facilitate collaboration.

6 Organizers

Prateek Yadav is a Ph.D. student at the University of North Carolina at Chapel Hill advised by Prof. Mohit Bansal. His is research focuses on continuous model adaptation and composition to enable decentralized collaborative development of models through modular architectures, efficient communication methods for updates, and merging technique. Specifically, he has previously worked on model merging [Yadav et al., 2024b,c], continual learning [Yadav and Bansal, 2023, Yadav et al., 2023b], mixture-of-expert models [Li et al., 2024, Li et al., Yadav et al., 2024a], training/inference efficiency [Yadav et al., 2023a], reasoning, and graph neural networks. He was an organizer of the NeurIPS 2024 LLM Merging Competition [Tam et al., 2024].

Haokun Liu is a Ph.D. student at the University of Toronto and Vector Institute, advised by Colin Raffel. His research aims to specialize general models to various downstream use cases and to transfer the knowledge in specialized models back to a general model or other specialized models. He received his M.Sc. from New York University and spent time at Allen Institute for AI, Google Translate, and MIT-IBM Waston AI Lab.

Wanru Zhao is a PhD candidate in the Department of Computer Science at the University of Cambridge, supervised by Prof. Nic Lane, and a visiting researcher at Vector Institute, working with Prof. Colin Raffel. Her research targets the development of more decentralized AI by focusing on algorithm and system co-design that are less centralized and more collaborative to achieve better performance and efficiency under data-sharing constraints. She has helped organize the Oxbridge Women in Computer Science Network and has been a speaker at Cambridge Women@Computer Lab Tech Talklets and Microsoft Research.

Arthur Douillard is a research scientist at Google DeepMind. His research focuses on continual learning and decentralized modular networks. In particular, he aims to build a model which would be split apart in modules across the world and multiple teams would each train a subset of that modular network. Arthur is the organizer of the Modularity annual workshop in Google DeepMind.

Marco Ciccone is a Postdoctoral Fellow at the Vector Institute working with Prof. Colin Raffel. His research focuses on building modular intelligent systems that can learn to solve new problems by reusing, composing, and improving previously learned skills. This includes research on Federated Learning to enable collaboration across users by sharing knowledge and resources while preserving privacy; Continual Learning, to enable continuous knowledge integration allowing models to keep improving without forgetting, and Modular Learning, for designing more reusable, maintainable, and efficient models. He served as Competition Track co-chair at NeurIPS in 2021, 2022 and 2023.

Colin Raffel is an associate professor of computer science at the University of Toronto, an associate research director at the Vector Institute, a faculty researcher at Hugging Face, and the president of the board of EleutherAI. His lab works on methods for decentralized and collaborative machine learning, efficient training recipes, and characterizing and mitigating risks associated with large-scale AI. He was an organizer of the NeurIPS 2024 LLM Merging Competition, the NeurIPS 2022 Workshop on Transfer Learning for NLP, the ICML 2022 Workshop on Pre-Training, the ICLR 2021 Workshop on Enormous Language Models, the NeurIPS 2020 Competition on Efficient Open-Domain Question Answering, and the late-breaking/demo session of ISMIR 2016. He has been a senior area chair at NAACL, EMNLP, NeurIPS, and ICLR. He also was the founder and lead organizer of the Hacking Audio and Music Research event series, with twelve hackathons organized internationally from 2013-2018.

References

Dana H Ballard. Cortical connections and parallel processing: Structure and function. *Behavioral and brain sciences*, 9(1):67–90, 1986.

Devvrit, Sneha Kudugunta, Aditya Kusupati, Tim Dettmers, Kaifeng Chen, Inderjit Dhillon, Yulia Tsvetkov, Hannaneh Hajishirzi, Sham Kakade, Ali Farhadi, and Prateek Jain. Matformer: Nested transformer for elastic inference, 2023. URL https://arxiv.org/abs/2310.07707.

Michael Diskin, Alexey Bukhtiyarov, Max Ryabinin, Lucile Saulnier, Quentin Lhoest, Anton Sinitsin, Dmitry Popov, Dmitry Pyrkin, Maxim Kashirin, Alexander Borzunov, Albert Villanova del Moral, Denis Mazur, Ilia Kobelev, Yacine Jernite, Thomas Wolf, and Gennady Pekhimenko. Distributed deep learning in open collaborations, 2021. URL https://arxiv.org/abs/2106.10207.

Arthur Douillard, Qixuan Feng, Andrei A. Rusu, Adhiguna Kuncoro, Yani Donchev, Rachita Chhaparia, Ionel Gog, Marc'Aurelio Ranzato, Jiajun Shen, and Arthur Szlam. Dipaco: Distributed path composition, 2024. URL https://arxiv.org/abs/2403.10616.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts, 2014. URL https://arxiv.org/abs/1312.4314.

Anastasiia Filippova, Angelos Katharopoulos, David Grangier, and Ronan Collobert. No need to talk: Asynchronous mixture of language models, 2024. URL https://arxiv.org/abs/2410.03529.

Jerry A Fodor. *The modularity of mind*. MIT press, 1983.

Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3 (4):128–135, 1999.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. Training compute-optimal large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=iBBcRUlOAPR.

Sami Jaghouar, Jack Min Ong, and Johannes Hagemann. Opendiloco: An open-source framework for globally distributed low-communication training, 2024. URL https://arxiv.org/abs/2407.07852.

Nikhil Kandpal, Brian Lester, Mohammed Muqeeth, Anisha Mascarenhas, Monty Evans, Vishal Baskaran, Tenghao Huang, Haokun Liu, and Colin Raffel. Git-theta: A git extension for collaborative development of machine learning models. 2023. URL https://arxiv.org/abs/2306.04529.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. Sparse upcycling: Training mixture-of-experts from dense checkpoints, 2023. URL https://arxiv.org/abs/2212.05055.

Pingzhi Li, Zhenyu Zhang, Prateek Yadav, Yi-Lin Sung, Yu Cheng, Mohit Bansal, and Tianlong Chen. Merge, then compress: Demystify efficient smoe with hints from its routing policy. In *The Twelfth International Conference on Learning Representations*.

Pingzhi Li, Prateek Yadav, Jaehong Yoon, Jie Peng, Yi-Lin Sung, Mohit Bansal, and Tianlong Chen. Glider: Global and local instruction-driven expert router, 2024. URL https://arxiv.org/abs/2410.07172.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data, 2023. URL https://arxiv.org/abs/1602.05629.

David Lorge Parnas. On the criteria to be used in decomposing systems into modules. *Communications of the ACM*, 15(12):1053–1058, 1972.

Jonas Pfeiffer, Sebastian Ruder, Ivan Vulić, and Edoardo Ponti. Modular deep learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=z9EkXfvxta. Survey Certification.

Colin Raffel. Building machine learning models like open source software. *Communications of the ACM*, 66(2):38–40, 2023.

Alexandre Ramé, Guillaume Couairon, Mustafa Shukor, Corentin Dancette, Jean-Baptiste Gaya, Laure Soulier, and Matthieu Cord. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards, 2023. URL https://arxiv.org/abs/2306.04488.

David Raposo, Sam Ritter, Blake Richards, Timothy Lillicrap, Peter Conway Humphreys, and Adam Santoro. Mixture-of-depths: Dynamically allocating compute in transformer-based language models, 2024. URL https://arxiv.org/abs/2404.02258.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017. URL https://arxiv.org/abs/1701.06538.

Sainbayar Sukhbaatar, Olga Golovneva, Vasu Sharma, Hu Xu, Xi Victoria Lin, Baptiste Rozière, Jacob Kahn, Daniel Li, Wen tau Yih, Jason Weston, and Xian Li. Branch-train-mix: Mixing expert llms into a mixture-of-experts llm, 2024. URL https://arxiv.org/abs/2403.07816.

Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38, 2019.

Derek Tam, Margaret Li, Prateek Yadav, Rickard Brüel Gabrielsson, Jiacheng Zhu, Kristjan Greenewald, Mikhail Yurochkin, Mohit Bansal, Colin Raffel, and Leshem Choshen. LLM merging: Building LLMs efficiently through merging. In *NeurIPS 2024 Competition Track*, 2024. URL https://openreview.net/forum?id=TiRQ4Gl4Ir.

Günter P. Wagner, Jason Mezey, and Raffaele Calabretta. Natural Selection and the Origin of Modules. In *Modularity: Understanding the Development and Evolution of Natural Complex Systems*. The MIT Press, 05 2005. ISBN 9780262269698. doi: 10.7551/mitpress/4734.003.0009. URL https://doi.org/10.7551/mitpress/4734.003.0009.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time, 2022. URL https://arxiv.org/abs/2203.05482.

Prateek Yadav and Mohit Bansal. Exclusive supermask subnetwork training for continual learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 569–587, 2023.

Prateek Yadav, Leshem Choshen, Colin Raffel, and Mohit Bansal. Compeft: Compression for communicating parameter efficient updates via sparsification and quantization, 2023a. URL https://arxiv.org/abs/2311.13171.

Prateek Yadav, Qing Sun, Hantian Ding, Xiaopeng Li, Dejiao Zhang, Ming Tan, Parminder Bhatia, Xiaofei Ma, Ramesh Nallapati, Murali Krishna Ramanathan, et al. Exploring continual learning for code generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 782–792, 2023b.

Prateek Yadav, Colin Raffel, Mohammed Muqeeth, Lucas Caccia, Haokun Liu, Tianlong Chen, Mohit Bansal, Leshem Choshen, and Alessandro Sordoni. A survey on model moerging: Recycling and routing among specialized experts for collaborative learning, 2024a. URL https://arxiv.org/abs/2408.07057.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36, 2024b.

Prateek Yadav, Tu Vu, Jonathan Lai, Alexandra Chronopoulou, Manaal Faruqui, Mohit Bansal, and Tsendsuren Munkhdalai. What matters for model merging at scale? *arXiv preprint arXiv:2410.03617*, 2024c.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew Dai, Zhifeng Chen, Quoc Le, and James Laudon. Mixture-of-experts with expert choice routing, 2022. URL https://arxiv.org/abs/2202.09368.