Open Source Gen AI

NYC

March 25-26
2024

WORKSHOP

NSF

Workshop Participants

Alexander Rush	Cornell Tech & Hugging Face
Colin Raffel	University of Toronto & Hugging Face
Danqi Chen	Princeton University
Daphne Ippolito	Carnegie Mellon University & Google Deepmind
Emma Strubell	Carnegie Mellon University & Allen Institute for AI
Eugene Cheah	RWKV OSS & Recursal AI
Graham Neubig	Carnegie Mellon University
Greg Leppert	Harvard The Berkman Klein Center, The Library Innovation Lab
Hanna Hajishirzi	University of Washington & Allen Institute for AI
Hao Zhang	UCSD
Irina Rish	Mila
John/Austin Cook	Alignment Lab AI
Leshem Choshen	MIT IBM
Louis Castricato	EleutherAI & Synth Labs
Luca Soldaini	Allen Institute for AI
Ludwig Schmidt	Anthropic, University of Washington, Stanford University
Niklas Muennighoff	Contextual AI & Allen Institute for AI
Peter Henderson	Princeton University
Sara Hooker	Cohere for AI
Soumith Chintala	Meta
Stella Biderman	EleutherAI
Susan Zhang	Google Deepmind

Swabha Swayamdipta	University of Southern California
Tatsunori Hashimoto	Stanford University
Tegan Maharaj	University of Toronto, Schwartz-Reisman Institute for Society and Technology, Vector, Mila
Teven Le Scao	Mistral
Tim Dettmers	University of Washington
Wenting Zhao	Cornell
Willie Neiswanger	University of Southern California
Yacine Jernite	Hugging Face
Yann LeCun	Meta & NYU
Ying Sheng	Stanford University
Yoav Artzi	Cornell Tech
Zhengzhong Liu	Carnegie Mellon, MBZUAI, Petuum

Table of Contents

Introduction

Generative AI is at a critical point in its growth trajectory. The last year has seen unprecedented excitement and usage, most notably with the commercialization of large language models (LLMs), as well as developments in other modalities such as generative image, speech and video models. Much of the effort, investment, and attention in this space has been concentrated in a few large centralized organizations, such as OpenAI, Google, and Microsoft, and their corresponding proprietary models.

The goal of this proposal is **to help foster a robust and distributed open-source ecosystem for generative AI that is comparable to the open-source software ecosystem**. We recognize that technological systems underlying Generative AI present novel and complex issues. Successes of Generative AI are not primarily code; they are the product of several factors: carefully coordinated data curation, strategically coordinated training runs, tuning with large-amounts of human feedback, and rigorous evaluation on realistic use-cases. As currently conceived, successful systems resemble a more monolithic, centralized development process than the decentralized open model. These challenges are socio-technical; we need both new research ideas as well as integration within open organizations.

In the face of these challenges, there is a thriving ecosystem of developers and researchers pushing forward open source AI. These groups take very different forms, from fully distributed collective organizations, to startups focusing on open-source tooling, to larger tech companies that have released critical models. In addition research groups in academia have specifically focused on key challenges for open source usage such as efficient fine-tuning, instruction tuning, model merging, and holistic evaluation, as well as a community that has begun to map out considerations for safe and equitable deployment of open-source models. With some exceptions, these groups primarily have operated independently on a variety of different topics.

To build on these efforts, Cornell Tech was funded by NSF to host a workshop on "**Open-Source Generative AI**" to bring together leaders in open-source from different organizations to gather input on the core challenges and opportunities in building a robust open-source foundation for generative AI. With this funding, Cornell Tech organized a two day workshop on this subject.

- Open-Source Generative AI Day 1 - Monday, March 25, 2024 9-5 ET
- Open-Source Generative AI Day 2 - Monday, March 26, 2024 9-4 ET

In this workshop the participants discussed the **central areas where open-source can effectively contribute to the development of generative AI and what are some of the ambitious ways organizations can expand their potential impact.** The participants still feel this is an early area of development, so there was not a push to arrive at specific long-term recommendations. However, many of the talks in the workshop discussed some of the core (non-intuitive) lessons from real-world open-source deployments of generative AI. This report will summarize the key lessons from the workshop and overarching themes.

The general question discussed throughout was "What are the key challenges in open-source development, training, and use of generative AI?" While this seems like a straightforward question, much of the discussion touched on how views on this topic have shifted, e.g. from a primary focus on training to data and analysis.

During the workshop participants pointed to **six themes for open-source generative AI development**. These included:

- Data Collection and Curation
- Distributed Human Feedback and Evaluation
- Organizational Structure for Generative AI
- Training Frontier Models
- Enabling Accessible User Models
- Model Safety

These areas overlap in key ways and many participants noted that they are all necessary for successful development of open-source Generative AI systems. Unlike some other aspects of the Generative AI pipeline they seem unlikely to be targetable by software systems alone. Therefore we believe prioritizing these goals can best facilitate sustainable and robust open-source generative AI systems and organizations. For each theme we consider the a) *context* in current generative AI, b) *opportunities* for open-source communities to engage, c) *challenges* in the development of this theme, and d) *next steps* for future work.

This document is not intended to be a consensus of the workshop, but a summarization of common thoughts and priorities. As the workshop was closed, we will not put names to individual comments or quotations. Points of disagreement will also be marked.

Theme 1: Data Collection and Curation

Context

It is necessary to have open generative models that are transparent in their training data in addition to weights. While closed-data, open-weight models are now numerous, these artifacts provide no recipe for replicability, no transparency as to their data sources, and no ability to retrain with variant data. In addition, few large datasets are documented and available, despite the centrality of data quality in model accuracy. Collecting and curating these datasets has been a critical process that several open-source organizations have put significant effort into curating.

Opportunities

Data for generative AI systems mostly comes from openly available sources and is accessible for the creation of open models. Unlike training, data processes can also be distributed to broad teams and does not require significant GPU resources. While these processes are still costly, they are on a scale that is more plausible for open organizations. Open organizations can explore better methods for curating data and make progress on smaller scales. Research participants discussed methods for learning better data filtering. Ongoing competitions, such as DataComp, look to improve the data curation process in reproducible ways. Past success in replicating models such as CLIP shows promise in open data efforts.

Challenges

Guidance on best practices for data collection is sparse. Collecting data is not removed from model training, and training large models requires training many smaller scale models to assess the quality of data, which requires significant compute. In addition better tools are needed for deduplicating, examining, and testing data collections. Generally methods for getting insights into large datasets are still nascent, and currently this process requires domain expertise.

Next Steps

A central research question is how pretraining data impacts the performance of large language models (LLMs). To answer this question, it is essential to understand the data both qualitatively and quantitatively. Several active efforts contributing to this area include: (1) developing open-source tools to both process and analyze data ranging from terabytes to petabytes, (2) working with data with different levels of license permissiveness, and (3) investigating how controlling pretraining data can steer the behavior of LLMs -- specifically, how different data mixtures influence various behaviors in LLMs.

Theme 2: Distributed Human Feedback and Eval

Context

A success of open-source development of LLMs has been improving the performance of open-weight base models with data from large proprietary models. Several of the workshop participants reflected on this process. These participants noted that despite the early interest in these models, the main challenge they faced in this process was not the development of the model, but getting data to further refine and evaluate these systems. There are several challenges in this process: the data needs to be high-quality, diverse, and often dynamic, i.e. using continuous human interaction from users to improve and stay current. In addition the systems for data collection must be interesting products that users want to interact with.

Opportunities

Distributed human feedback for generative AI is a major opportunity for open-source. Collecting human feedback for real generative AI systems can be done in a continually improving manner. Building communities where users actively provide evaluations and feedback on generative AI systems, is something open-source communities are already doing well and has precedent for scaling. Open systems are beginning to be used for dynamic human evaluation of many generative AI systems and are thought to be trustworthy.

Challenges

There are significant barriers to open-source collection of human feedback. One issue is the marginal cost of data collection from open-source systems. In addition to instrumenting open-source tools to collect user feedback, this data needs to be stored and ideally shared to reach the scale necessary to build instruct-like models. A related issue is that open-source users value privacy and would be unlikely to use systems that collect feedback. Finally, determining how to effectively incentivize users to provide feedback remains a significant challenge. For example, it was observed that fewer than 10% of users who interact with chatbots left any feedback, even when it involved just a simple voting between responses from two chatbots.

Non-english LLMs also pose a major challenge for evaluations, with one participant noting the "multilinguality is a cliff" for evaluation. Challenges for multilingual evaluation include where to get data for various languages, how to carefully avoid contamination, and how to pick what is being measured. These issues are further complicated by the range of organizations doing evals and the difficulty of ensuring reproducibility.

Next Steps

Evaluation is an area open-source can continue to play a leading role. One advantage is that open-source evals can avoid doing it "how GPT does it". A key goal is to produce evals that match how models are being used by users in their own language. Dynamic evaluation from real users is a clear goal as well, and producing systems that get consistent feedback from a range of users on active language models is necessary. This is particularly needed in domains (like code) where you may need specialist participation.

Theme 3: Organizational Structures for Generative AI

Context

Open-source communities require organizational structures that can support the responsibilities and issues that arise from data curation and community management. There now exist several different large communities that have proved the ability to produce Generative AI systems without institutional affiliation. One participant noted that "AI has broken free from academia". Much of this work is being done through Discord, with some organizations hosting thousands of participants in their channels. While participants noted this as a success of open-systems, there were also questions about the sustainability of these processes, and particularly some of the notable issues already seen with open organizations and data.

Opportunities

One theme raised by several participants was the relationship of large scale data curation to organizations like libraries and museums that have developed processes for data curation and management. Generally there was a sense that there had not been enough interaction between Generative AI and other expert archivists. There is an opportunity for these organizations to improve the traceability of data sources and their accessibility. One participant expressed a hope that organizations could move from "an anarchist mafia of vigilantes" to storers of knowledge.

Challenges

While several Gen AI groups are run autonomously, there are notable points of centralization, specifically the reliance on large sources of compute from groups that may not be sustainable. In addition, GenAI organizations seem to struggle with the ability to divide big projects into smaller components, which is thought necessary for open-source organizational success. Another perennial question in open-source is the economic incentives that participants have. Much of the current work in development and evaluation is done by unpaid enthusiasts, which has led to unpredictable trends. For instance some of the evaluation infrastructure relies on volunteer participants, which makes it hard to evaluate less popular tasks or less used languages.

There is also the overarching question of how AI fits into the general definition of open-source. Open-weight models, without code or data, have been undoubtedly useful but are outside of the traditional open-source categories. Other models may or may not fall in the categories defined by the OSI. In addition, training data and synthetic generated data present novel challenges in both the legal and open-source category.

Finally, the workshop discussed how the current emphasis on the credit assignment mechanism in academia does not necessarily reward researchers for building open-source AI. This is because such projects typically require a large team working collaboratively without highlighting the contributions of individual researchers. This particularly disadvantages junior researchers who participate in long-term, highly impactful open-source projects.

Next Steps

A major question that arose is "what would the Linux of generative AI look like (and is that desirable)". The BLOOM project was one effort along these lines, but it was a one time effort. If started again, would need long term compute and agreed upon metrics. Major advantage would be distributed community eval. Universities provide one method for pooling compute, but have infrastructure challenges in terms of data centers not designed for efficient training. The standard open-source model is also held back by the lack of analogues to forking and merging, which are being explored in the open-source and research community. Need a technical mechanism for projects to branch off and continue.

Theme 4: Training Frontier Models

Context

"Language model training is as much an art as a science." Yet it is still important for this art to be known and developed in open and accessible communities. Both academia and open communities have struggled to keep up with the computational needs required to train models at the frontiers of GenAI. However, several participants have been able to train successful generative AI models and the necessary skill and knowledge are becoming well documented in open-organizations.

Opportunities

Significant computational resources exist for training large models outside of the "standard" training setup of NVidia GPUs or Google TPUs. Several participants discussed experiences training large generative models using AMD, Cerberas, or Intel architectures in a diverse set of different computing environments. While there were additional challenges in getting these environments to be as effective as more battle hardened environments, there are significant benefits.

Other participants encouraged open-source to embrace the resource challenges and to develop better approaches to training and scaling models. Ideas include more effective data mixtures, pruning and expansion of models, model merging and reuse, and distributed architecture search.

Challenges

The challenges of creating an open-source pretrained model at the frontier have been widely discussed. In addition to compute and training details, there are questions of access to proprietary data, the inclusion of synthetic data, knowledge of training details (such as hyperparameters), and the post-training data and expertise. There was also disagreement about whether distributed training presents more of a technical challenge or a social challenge.

Next Steps

Several breakout groups discussed the challenge of building an open source LLMs at the level of GPT-4. There were several key questions about how this might proceed. a) would it need to be centralized or could it be done in a distributed manner across multiple locations or models. b) could continual training be an area that open source could improve upon, c) could open source be an area where non-standard models or training styles could be explored.

Theme 5: Accessible User Models

Context

Primary focus of proprietary models has been bigger and more powerful; open needs to be about systems that a broad community of people can actually use. Open-weight models have improved on this front, but they are still very large and challenging to use. Prompting can be an effective and accessible method for model usage but can also be challenging for users to employ effectively and has turned users away from generative AI.

Opportunities

A straightforward opportunity for open-source generative AI is to produce smaller and more compute accessible models. Making more accessible models will expand not just the use but also the inclusive aspects of participants in open AI projects. Several participants noted that techniques such as model merging or model routing present a technical path forward to have less compute requirements for open-source generative AI.

A more ambitious project is the development of decentralized inference and fine-tuning capability. While there are existing projects for distributed inference, training remains a challenging technical problems and participants disagreed about the feasibility of practical distributed training in the short term.

Challenges

A subtheme of this area is that Generative AI research now resembles a "research funnel" where major systems require huge upfront capital investment. This looks almost more like drug development than traditional AI. However, continuing the analogy, a challenge here is that we lack a mouse model that can be used to find promising areas of study at the smaller scale.

Next Steps

A breakout session discussed "Scaling Down Models and Data". A major question is understanding what scales are useful for predicting large scale performance. For instance the 0.5B parameter scale is now accessible to most experimenters but it is unclear whether results at this scale are generalizable. In addition the answer may be different for data experiments vs architecture experiments. Collecting and validating these experiments across open-organizations is a critical challenge.

Theme 6: Model Safety

Context

Open-source LLMs systems have come under scrutiny recently due to some of their safety concerns. For example there are instances where models can provide instructions for self-harm or general non-consensual pornography. There is a non-zero chance these models may come under export control or similar restrictions. Some participants felt that this concern was over-stated, particularly widely publicized concerns over the risk of access to bioweapon information. Others noted a need to actively study technical methodology to reduce this risk.

Opportunities

Open models provide a mechanism for studying model safety in a reproducible and testable manner. Open organizations provide artifacts and expertise to enable better interpretability and alignment research. Research was presented that showed how open-models allow the exploration of models that can "self-destruct" if attacked or be used to study and model certain attacks. Other work looked at how open models could be used to examine closed-source models, for example to estimate model parameters or sizes.

Safety is also closely connected to data. Open data collections can be used to collect data for toxicity evaluation, or to provide shared tools to allow open tools for data filtering and curation, e.g. a shared directory of harmful sites.

Challenges

So far there has not been much success with technical solutions to model safety concerns. There are three major sources of challenges: (1) how to balance performance with safety, given that models designed with enhanced safety often exhibit reduced utility. (2) The concept of safety is broad, and there is still no consensus on how to define and categorize its different aspects. For example, determining toxic content can be culturally specific. (3) The lack of effective techniques to improve the safety of models. Safety finetuning does not seem to make (open or closed) models more safe and can be easily bypassed. While there are promising methods for stronger safety, they are challenging to scale to larger models. Filtering data can prevent certain issues, but there is also evidence that removing harmful data from training can make models worse at detection of harmful text. Open-weight models also seem to be poorly aligned to user intention and can be worse than proprietary models in their controlled responses.

Next Steps

Several participants noted that open models are seen as "non-safe" but that it was important for open-source organizations to work towards rebutting this perception. Factors like open data, transparent evaluation and reporting, and technical work into model guardrails would improve that perception, and that open models could be seen as significantly safer than closed systems.