AI / OPEN SOURCE

How Important Is Open Source to AI Adoption?

Open source projects are critical to developers of AI and ML projects; we examine usage statistics from a number of different studies.

Aug 23rd, 2023 7:36am by Lawrence E Hecht

Image via Unsplash.

I HAVE AN OPINION

We'd love to hear what you think.

How important is open source to the future of AI and LLMs? The answer depends on how you define open source in the age of AI.

Open source is viewed by 40% as "the solution" to concerns about AI ownership, while only 15% disagree with this assessment, according to the 224 UK residents surveyed for State of Open: The UK in 2023, Phase Two. In this regard, many respondents are talking about who should own the large datasets being generated by large language models (LLMs).

Indeed, Predibase's just published "Beyond the Buzz: A Look at Large Language Models in Production" found there is a reluctance to rely on commercial LLMs in production. Based on a survey of 150 people conducted from May through July 2023, 13% of respondents say their enterprise has at least one LLM in production. Another 44% said their organization has so far only used LLM for experimentation purposes.

Among the whopping 85% of survey respondents who are using or planning to use LLMs, only 27% actually expect a commercial version to be used in production. Almost half (49%) of those with no plans to use a commercial LLM cited a desire not to share proprietary information with vendors. In comparison, only 17% said the reason is because commercial LLMs are too expensive to scale.

Stagnating Growth in Open Source AI

Despite all the noise surrounding the subject, the growth in new traditional AI projects continues to slow down. According to the OECD

7% by 2022.

With the boom in LLMs and applications that take advantage of them, it is likely that the number of AI projects is being under-counted because AI-related concepts identified have changed over time. Indeed, Ashley Wolf, Open Source Program Office Director at GitHub, told The New Stack that "it is possible some projects may have transitioned to using new terminology that isn't currently reflected. Additionally, there might be a trend where people are focusing on highly successful projects, resulting in less churn. Both are worth investigating."

Contributions to Open Source AI

Open source projects are obviously critical to developers of artificial intelligence and machine learning projects. Eighty-nine percent (89%) of all developers involved with AI/ML have contributed to an AI project, according to the AI & Machine Learning Survey Report, which was published by Evans Data in Q2 2023. That statistic is comparable to the results published by SlashData in Q3 2022, in which 73% of all developers contribute to a "vendor-owned" open source community.

Both these stats likely vastly overstate the act of using an open source project without actually contributing to one, according to the JetBrains State of Developer Ecosystem 2022. That study found that of 424 developers involved with machine learning activities, only 54% have contributed — with almost half (45%) of these contributors having only contributed a few times in their careers. It is likely that some people who claim to be contributors in the other studies are users, not contributors to projects.

Which open source communities' AI projects do you contribute to most often?

Python Software Foundation	44%
Apache Software Foundation	36%
PyTorch community	35%
TensorFlow community	34%
Linux Foundation	31%
Vendor-run open source communities	22%
Open Neural Network Exchange (ONNX)	20%
Other	2%
None – I do not contribute	11%

There continues to be confusion about what exactly makes a community vendor-dominated. Using something called an Elephant Factor, vendor control can be defined based on how many companies account for at least half of all contributions. A project can also be considered to be vendor-0wned if it is located in a repository controlled by a corporate organization (i.e., anything hosted in https://github.com/openai). Another approach is to look at the governing structure of a project.

Per Evans Data, Python and Apache communities associated with AI are most likely to receive contributions from AI/ML developers. PyTorch supposedly receives contributions from 35% of AI/ML developers, with TensorFlow close behind at 34%. While the Meta donated PyTorch is now controlled by an independent foundation, Google still manages TensorFlow's progress. Even excluding these types of communities, 22% of AI/ML developers contribute to vendor-run communities.

Adoption of AI Frameworks

not exactly comparable, 71% use Scikit-Learn, which is a Python library for machine learning that utilizes several other popular Python projects. The use of these frameworks is not widespread among all professional developers, however, with adoption rates between 9% and 10%.

25% of all AI/ML developers in the Stack Overflow survey have used Hugging Face Transformers (founded in 2019) extensively in the last year. Used by 21% of AI/ML developers, Nvidia's CUDA is the only non-open source framework on this list. CUDA allows software to use certain types of graphics processing units (GPUs).

54% of ML or Data Science Specialists Utilize PyTorch

Which other frameworks and libraries have you done extensive development work in over the past year?*

*N=1,510 data scientist or machine learning specialists. *The chart does not display the data for several frameworks not directly involved with machine learning.*

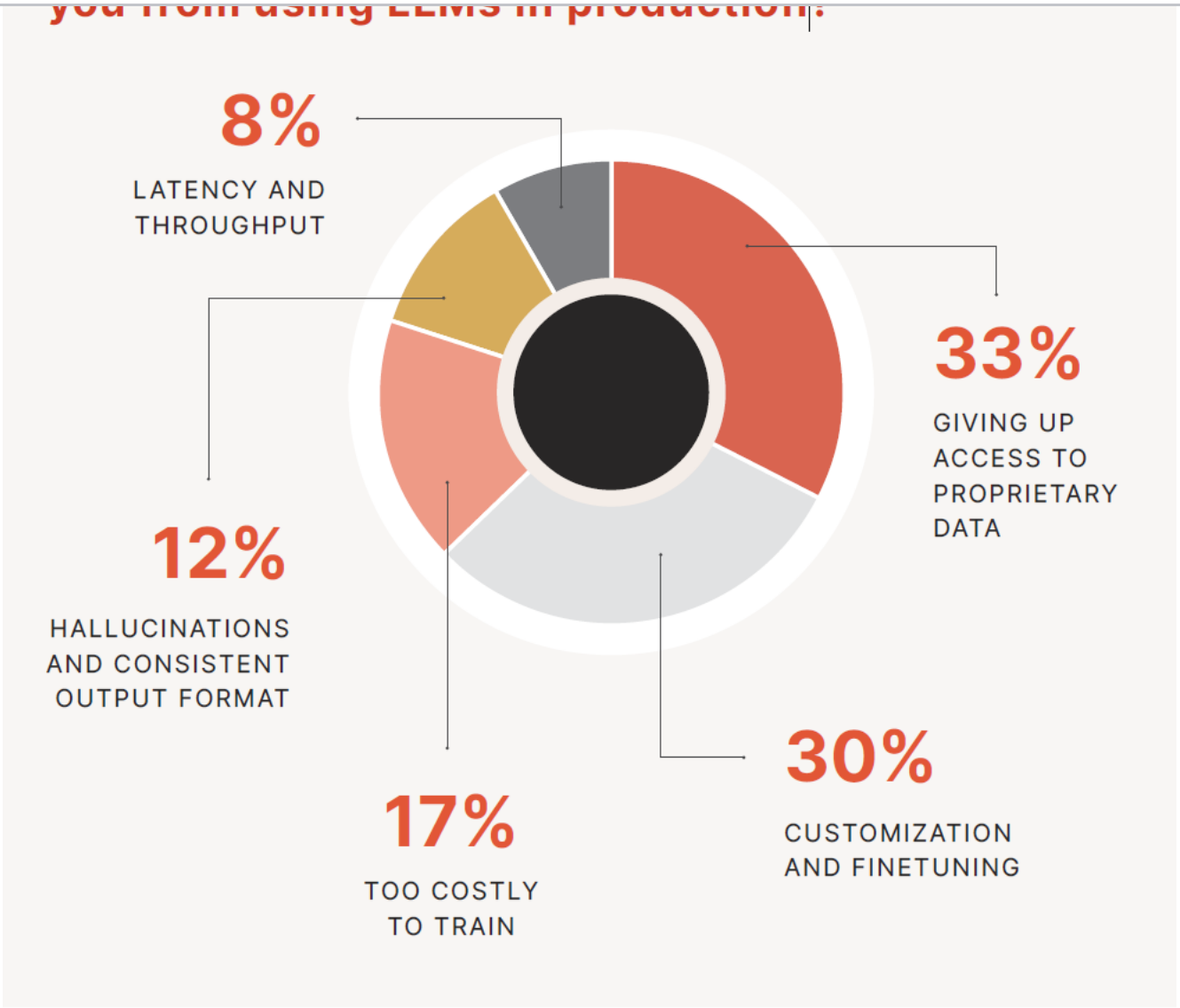
GPUs and the Edge

According to the aforementioned Evans Data report, 57% of AI/ML developers prefer using GPUs "that are dedicated to my individual development work" as compared to 42% who would rather share GPUs across multiple devs or workloads. The preference for using dedicated

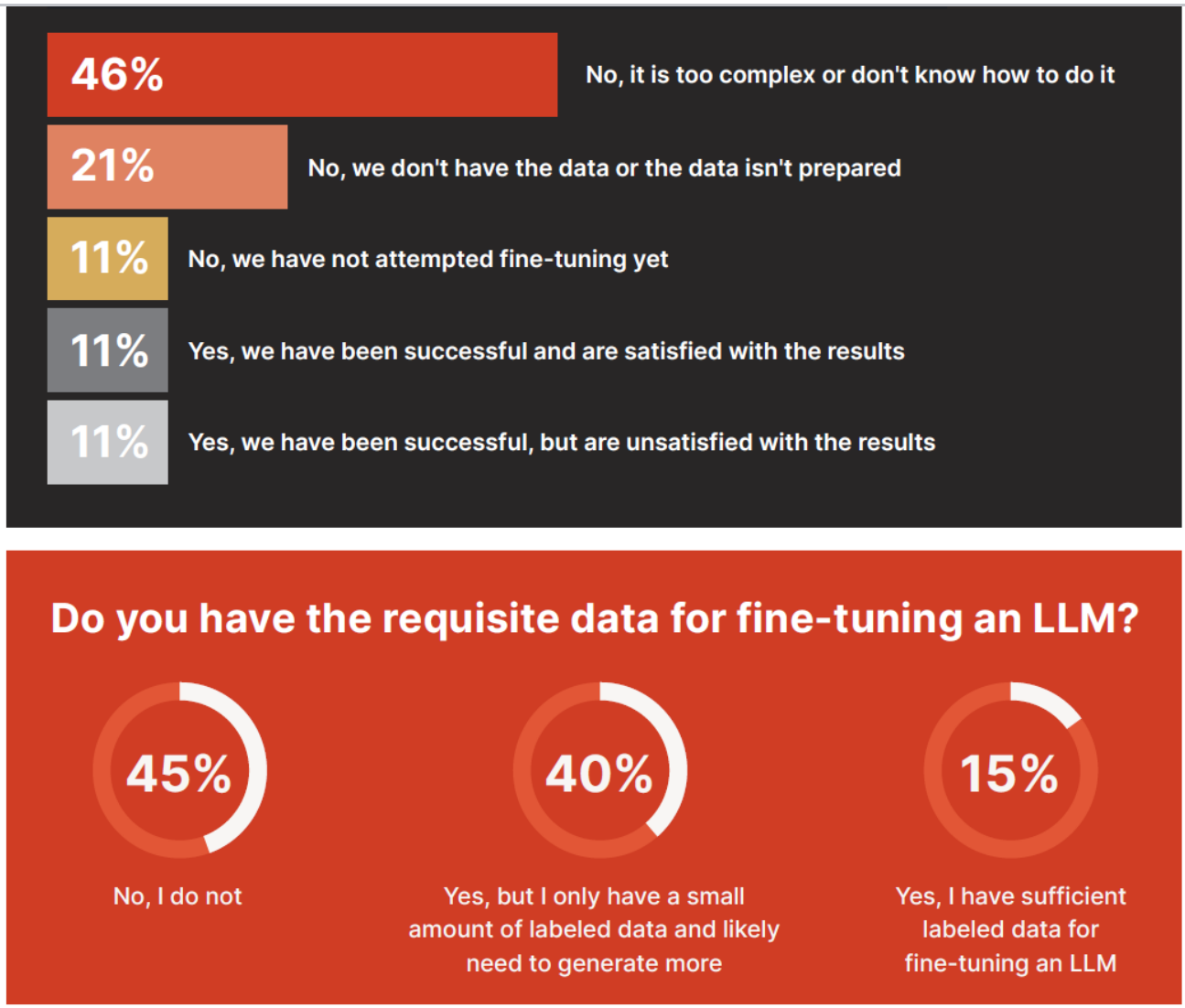
the proliferation of large language models (LLMs). Overall, 55% of AI/ML developers are running inference models at the edge, but these are most often (73% of the time) being deployed to run on PCs. As developers scale up their use of these models, they are expected to rely more heavily on specialized chips, which may often be installed by data center operators or cloud providers.

More from the Predibase Study

- Giving up access to proprietary data was cited by 33% of respondents as the top challenge preventing them from using LLMs in production. Customization and fine-tuning was the second most inhibitor to LLM use, cited by 30%.
- Digging into the challenge of fine-tuning a LLM, only 22% of the study have had success doing so. The top reason for not doing fine-tuning is they don't have the requisite knowledge to handle this complex task. In response to a different question, 45% said they do not have the data needed to fine-tune a LLM.



Source: Predibase's "Beyond the Buzz: A Look at Large Language Models in Production"



Source: Predibase's "Beyond the Buzz: A Look at Large Language Models in Production"

TNS

Lawrence has generated actionable insights and reports about enterprise IT B2B markets and technology policy issues for almost 25 years. He regularly works with clients to develop and analyze studies about open source ecosystems. In addition to his consulting work,...

Read more from Lawrence E Hecht →

TNS owner Insight Partners is an investor in: The New Stack.

SHARE THIS STORY

Unleash the Power of Generative AI to Shift Left

How Vector Search Can Influence Customer Shopping Habits

Cockroach Labs Chief Targets LLMs with Vector Encoding

Top 5 Large Language Models and How to Use Them Effectively

INSIGHTS FROM OUR SPONSORS

From Chaos to Consistency: A Comprehensive Approach to Maintaining a Drift-Free Infrastructure
17 August 2023

K8s PostgreSQL Operator
27 July 2023

Dynamic Cloud Interoperability: Redefining Cloud Agnosticism
2 June 2023

TensorFlow Lite Tutorial: How to Get Up and Running
23 August 2023

Telegraf Deployment Strategies with Docker Compose
21 August 2023

Choosing a Client Library When Developing with

Introducing ScyllaDB Enterprise 2023.1 – with Raft
22 August 2023

How Pinhome Improved Recommendation Engine Latencies by Moving from MongoDB and PostgreSQL to ScyllaDB
21 August 2023

5 Intriguing ScyllaDB Capabilities You Might Have Overlooked
14 August 2023

ARCHITECTURE

Cloud Native Ecosystem
Containers
Edge Computing
Microservices
Networking
Serverless
Storage

ENGINEERING

AI
Frontend Development
Software Development
Typescript
WebAssembly
Cloud Services
Data
Security

OPERATIONS

Platform Engineering
Operations
CI/CD
Tech Life
DevOps
Kubernetes
Observability
Service Mesh

CHANNELS

Podcasts

THE NEW STACK

About / Contact
Sponsors
Sponsorship
Contributions

FOLLOW TNS