Open source isn't ready for generative AI. How stakeholders are changing this light bulb together

Open-source licenses, already stretched thin by software-as-a-service and the cloud, are an even worse fit for AI's large language models. What's an open source leader to do?

Written by **Steven Vaughan-Nichols,** Senior Contributing Editor on Oct. 9, 2023

choness/Getty Images

Without open source, there is no AI. It's that simple. But, those same licenses have been showing their age: The Gnu General Public License (GPL), Apache License, and Mozilla Public License don't fit well with software-as-a-service or cloud services. AI poses even larger problems. The open-source licenses, with their copyright law foundations, aren't a good fit for AI's large language models (LLM)s.

This isn't just some theoretical techno-legal issue, either. It's already showing up in the courts.

Also: Open source is actually the cradle of artificial intelligence. Here's why

In J. Doe 1 et al. vs GitHub, the plaintiffs allege that Microsoft, OpenAI, and GitHub -- via their commercial AI-based system, OpenAI's Codex and GitHub's Copilot -- stole their open source code. The class action suit claims that code "suggested" by AI often consists of near-identical strings of code scraped from public GitHub repositories -- but without the required open-source license attributions.

On a related issue, two groups of writers, including George R.R. Martin, Michael Chabon, and John Grisham, are suing Microsoft and OpenAI for taking their work and using it in their LLMs. Copyright, the legal foundation of open source, is at the heart of this issue.

But this isn't simply a Microsoft problem.

As Sean O'Brien, Yale Law School lecturer in cybersecurity and founder of the Yale Privacy Lab, told my ZDNET colleague David Gewirtz: "I believe there will soon be an entire sub-industry of trolling that mirrors patent trolls, but this time surrounding AI-generated works. A feedback loop is created as more authors use AI-powered tools to ship code under proprietary licenses. Software ecosystems will be polluted with proprietary code that will be the subject of cease-and-desist claims by enterprising firms."

Others, like German researcher and politician Felix Reda, claim that all AI-produced code is public domain.

US attorney Richard Santalesa, a founding member of the SmartEdgeLaw Group, told Gewirtz that there exist both contract and copyright law issues -- and they're not the same thing. Santalesa believes companies producing AI-generated code will "as with all of their other IP, deem their provided materials – including AI-generated code – as their property." Besides, public domain code is not the same thing as open-source code.

Also: Red Hat's new rule: Open source betrayal?

So, what's to be done? Simply claiming your AI is open source is a nonstarter. Meta, for example, claims Llama 2 is open source. It's not.

As Erica Brescia, a managing director at RedPoint, the open source-friendly venture capital firm, asked on Twitter: "Can someone please explain to me how Meta and Microsoft can justify calling Llama 2 open source if it doesn't actually use an OSI [Open Source Initiative]-approved license or comply with the OSD [Open Source Definition]? Are they intentionally challenging the definition of OSS [Open Source Software]?"

Here's the short explanation: Meta is using open source as a marketing term, not a legal one. That usage won't fly once the lawsuits mount up

The problem with Llama 2 specifically is that it blocks extremely profitable companies from using it. According to Stephen O'Grady, open-source licensing expert and RedMonk co-founder, the problem is that they won't work in open source. "Imagine if Linux was open source unless you worked at Facebook,"

Also: Red Hat's new rule: Open source betrayal?

At the same time, as OpenUK CEO Amanda Brock observed, "I don't think we're going to see going forward any LLM or any significant AI being able to be licensed as open source, because the key to open source is the Open Source Definition."

And the road to that Definition was a long and bumpy one.

The first free software licenses began In the early 1980s when MIT Lab programmer Richard M. Stallman couldn't get an early laser printer, the Xerox 9700, to produce error messages. The problem? Stallman couldn't read or change its source code. At the time, this was a new development. Although we now think of proprietary software as the default, it wasn't then.

So, Stallman created the GNU General Public License (GPL). While not the *first* Free Software license (that honor belongs to the Berkeley Software Distribution (BSD) license), GNU would prove to be very influential. In no small part, that's because Linus Torvalds chose to use the GPLv2 as Linux's license.

The GPL is based on two principles. First, software code can be copyrighted. Second, anyone is free to read and edit the code so long as these freedoms aren't taken away from anyone else.

Also: A look back at 40 Years of GNU and the Free Software Foundation

By 1985, Free Software was becoming popular, but it also had become clear that the word "free" was too ambiguous. After Netscape released Mozilla's source code -- which became the basis of the Firefox web browser -- several leading Free Software luminaries, including Eric S. Raymond, Bruce Perens, Michael Tiemann, Jon "Maddog" Hall, and Christine Peterson, coined the phrase open source to describe this kind of license. In 1998, Perens and Raymond went on to found the OSI, which drafted the Open Source Definition (OSD) and used this as the general guide to defining all open-source licenses.

All open-source licenses must comply with the OSD. For AI and LLMs, that's much easier said than done.

True, there are open LLMs such as Falcon, FastChat-T5, and OpenLLaMA. But most LLMs contain proprietary, copyrighted, or simply unknown information that their owners won't tell you about. The Electronic Frontier Foundation (EFF) says it well: "Garbage In, Gospel Out."

We've seen this problem coming for a while. At Open Source Europe in Bilbao, Spain, last month, I spoke with Stefano Maffulli, executive director of the Open Source Initiative (OSI), the organization that defines and manages open-source licenses. "The process started two years ago when GitHub Copilot came out," Maffulli told me. "It was a watershed moment. All of a sudden, code you wrote as a human for humans, everything we have produced and put on the Internet was being harvested for machine learning."

Also: The best AI chatbots: ChatGPT and alternatives

So, what can we do? Maffulli and other open-source and AI leaders are working on combining AI with open-source licenses in sensible ways.

Maffulli observed that combining AI with open-source licenses is as hard, if not harder, than when software copyright was first applied to source code in the 1980s (when Free Software and open-source were first defined). True, open-source AI programs -- such as **TensorFlow**, **PyTorch**, and **Hugging Face** -- work well with old-style licenses. But old-style software isn't the problem. It's where software and data mix that the existing open-source licenses begin to break down. Specifically, it's where all that data and code merge together in AI/ML artifacts -- such as datasets, models, and weights -- that's where trouble emerges. "Therefore," said Mafulli, "we need to make a new definition for open-source AI."

This must be a definition that all stakeholders can agree upon and work with. Free software and open source are no longer just matters for developers. The goals of open-source savvy programmers and lawyers aren't the same as those of AI companies. To address this, Maffulli, together with Google, Microsoft, GitHub, Open Forum Europe, Creative Commons, Wikimedia Foundation, Hugging Face, GitHub, the Linux Foundation, ACLU Mozilla, and the Internet Archive, are working on a draft for defining a common understanding of open-source AI. In other words, all the AI players are working on the definition.

If all goes well, we can expect to see the fruits of their labor as early as this month. And while this will only be the first draft of the AI Open Source Definition, I expect that it will be finalized as quickly as possible. Everyone involved knows that AI is advancing rapidly and the sooner we get an open-source framework around it, the better.

/ artificial intelligence

The impact of artificial intelligence on software development? Still unclear

Android 14's AI-generated wallpapers are super fun. Here's how to create the

Editorial standards

show comments ↓

Part of a ZDNET Special Feature: The Intersection of Generative AI and Engineering

Home / Innovation / Artificial Intelligence

Open source is actually the cradle of artificial intelligence. Here's why

In the wildly competitive business of AI, is open source doomed to be always a bridesmaid, never a bride? Think again.

Written by **Steven Vaughan-Nichols,** Senior Contributing Editor on Oct. 9, 2023

Stabiilty.ai + Lightning.ai

In a way, open source and artificial intelligence were born together.

Back in 1971, if you'd mentioned AI to most people, they might have thought of Isaac Asimov's Three Laws of Robotics. However, AI was already a real subject that year at MIT, where Richard M. Stallman (RMS) joined MIT's Artificial Intelligence Lab.

/ special feature

The Intersection of Generative AI and Engineering

The surge of generative AI can harness tremendous potential for the engineering realm. It can also come with its challenges, as enterprises and engineers alike figure out the...

Read now →

Years later, as proprietary software sprang up, RMS developed the radical idea of Free Software. Decades later, this concept, transformed into open source, would become the birthplace of modern AI.

It wasn't a science-fiction writer but a computer scientist, Alan Turing, who started the modern AI movement. Turing's 1950 paper Computing Machine and Intelligence originated the Turing Test. The test, in brief, states that if a machine can fool you into thinking that you're talking with a human being, it's intelligent.

According to some people, today's AIs can already do this. I don't agree, but we're clearly on our way.

Also: The best AI chatbots: ChatGPT and alternatives

In 1960, computer scientist John McCarthy coined the term "artificial intelligence" and, along the way, created the Lisp language. McCarthy's achievement, as computer scientist Paul Graham put it, "did for programming something like what Euclid did for geometry. He showed how, given a handful of simple operators and a notation for functions, you can build a whole programming language."

Lisp, in which data and code are mixed, became AI's first language. It was also RMS's first programming love.

So, why didn't we have a GNU-ChatGPT in the 1980s? There are many theories. The one I prefer is that early AI had the right ideas in the wrong decade. The hardware wasn't up to the challenge. Other essential elements -- like Big Data -- weren't yet available to help real AI get underway. Open-source projects such as Hdoop, Spark, and Cassandra provided the tools that AI and machine learning needed for storing and processing large amounts of data on clusters of machines. Without this data and quick access to it, Large Language Models (LLMs) couldn't work.

Also: My two favorite ChatGPT Plus plugins and the remarkable things I can do with them

Today, even Bill Gates -- no fan of open source -- admits that open-source-based AI is the biggest thing since he was introduced to the idea of a graphical user interface (GUI) in 1980. From that GUI idea, you may recall, Gates built a little program called Windows.

In particular, today's wildly popular AI generative models, such as ChatGPT and Llama 2, sprang from open-source origins. That's not to say ChatGPT, Llama 2, or DALL-E are open source. They're not.

Oh, they were supposed to be. As Elon Musk, an early OpenAI investor, said: "OpenAI was created as an open source (which is why I named it "Open" AI), non-profit company to serve as a counterweight to Google, but now it has become a closed source, maximum-profit company effectively controlled by Microsoft. Not what I intended at all."

Also: The best AI image generators to try

Be that as it may, OpenAI and all the other generative AI programs are built on open-source foundations. In particular, Hugging Face's Transformer is the top open-source library for building today's machine learning (ML) models. Funny name and all, it provides pre-trained models, architectures, and tools for natural language processing tasks. This enables developers to build upon existing models and fine-tune them for specific use cases. In particular, ChatGPT relies on Hugging Face's library for its GPT LLMs. Without Transformer, there's no ChatGPT.

In addition, TensorFlow and PyTorch, developed by Google and Facebook, respectively, fueled ChatGPT. These Python frameworks provide essential tools and libraries for building and training deep learning models. Needless to say, other open-source AI/ML programs are built on top of them. For example, Keras, a high-level TensorFlow API, is often used by developers without deep learning backgrounds to build neural networks.

Also: Want to build your own AI chatbot? Say hello to open-source HuggingChat

You can argue until you're blue in the face as to which one is better -- and AI programmers do -- but both TensorFlow and PyTorch are used in multiple projects. Behind the scenes of your favorite AI chatbot is a mix of different open-source projects.

Some top-level programs, such as Meta's Llama-2, claim that they're open source. They're not. Although many open-source programmers have turned to Llama because it's about as open-source friendly as any of the large AI programs, when push comes to shove, Llama-2 isn't open source. True, you can download it and use it. With model weights and starting code for the pre-trained model and conversational fine-tuned versions, it's easy to build Llama-powered applications. There's only one tiny problem buried in the licensing: If your program is wildly successful and you have

> **greater than 700 million monthly active users in the preceding calendar month, you must request a license from Meta, which Meta may grant to you in its sole discretion, and you are not authorized to exercise any of the rights under this Agreement unless or until Meta otherwise expressly grants you such rights.**

You can give up any dreams you might have of becoming a billionaire by writing Virtual Girl/Boy Friend based on Llama. Mark Zuckerberg will thank you for helping him to another few billion.

Also: AI is a lot like streaming. The add-ons add up fast

Now, there do exist some true open-source LLMs -- such as Falcon180B. However, nearly all the major commercial LLMs aren't properly open source. Mind you, all the major LLMs were trained on open data. For instance, GPT-4 and most other large LLMs get some of their data from CommonCrawl, a text archive that contains petabytes of data crawled from the web. If you've written something on a public site -- a birthday wish on Facebook, a Reddit comment on Linux, a Wikipedia mention, or a book on Archives.org -- if it was written in HTML, chances are your data is in there somewhere.

So, is open source doomed to be always a bridesmaid, never a bride in the AI business? Not so fast.

In a leaked internal Google document, a Google AI engineer wrote, "The uncomfortable truth is, we aren't positioned to win this [Generative AI] arms race, and neither is OpenAI. While we've been squabbling, a third faction has been quietly eating our lunch."

That third player? The open-source community.

As it turns out, you don't need hyperscale clouds or thousands of high-end GPUs to get useful answers out of generative AI. In fact, you can run LLMs on a smartphone: People are running foundation models on a Pixel 6 at five LLM tokens per second. You can also finetune a personalized AI on your laptop in an evening. When you can "personalize a language model in a few hours on consumer hardware," the engineer noted, "[it's] a big deal." That's for sure.

Also: The ethics of generative AI: How we can harness this powerful technology

Thanks to fine-tuning mechanisms, such as the Hugging Face open-source low-rank adaptation (LoRA), you can perform model fine-tuning for a fraction of the cost and time of other methods. How much of a fraction? How does personalizing a language model in a few hours on consumer hardware sound to you?

The Google developer added:

> **"Part of what makes LoRA so effective is that -- like other forms of fine-tuning -- it's stackable. Improvements like instruction tuning can be applied and then leveraged as other contributors add on dialogue, or reasoning, or tool use. While the individual fine tunings are low rank, their sum need not be, allowing full-rank updates to the model to accumulate over time. This means that as new and better datasets and tasks become available, the model can be cheaply kept up to date without ever having to pay the cost of a full run."**

Our mystery programmer concluded, "Directly competing with open source is a losing proposition.... We should not expect to be able to catch up. The modern internet runs on open source for a reason. Open source has some significant advantages that we cannot replicate."

Also: Extending ChatGPT: Can AI chatbot plugins really change the game?

Thirty years ago, no one dreamed that an open-source operating system could ever usurp proprietary systems like Unix and Windows. Perhaps it will take a lot less than three decades for a truly open, soup-to-nuts AI program to overwhelm the semi-proprietary programs we're using today.

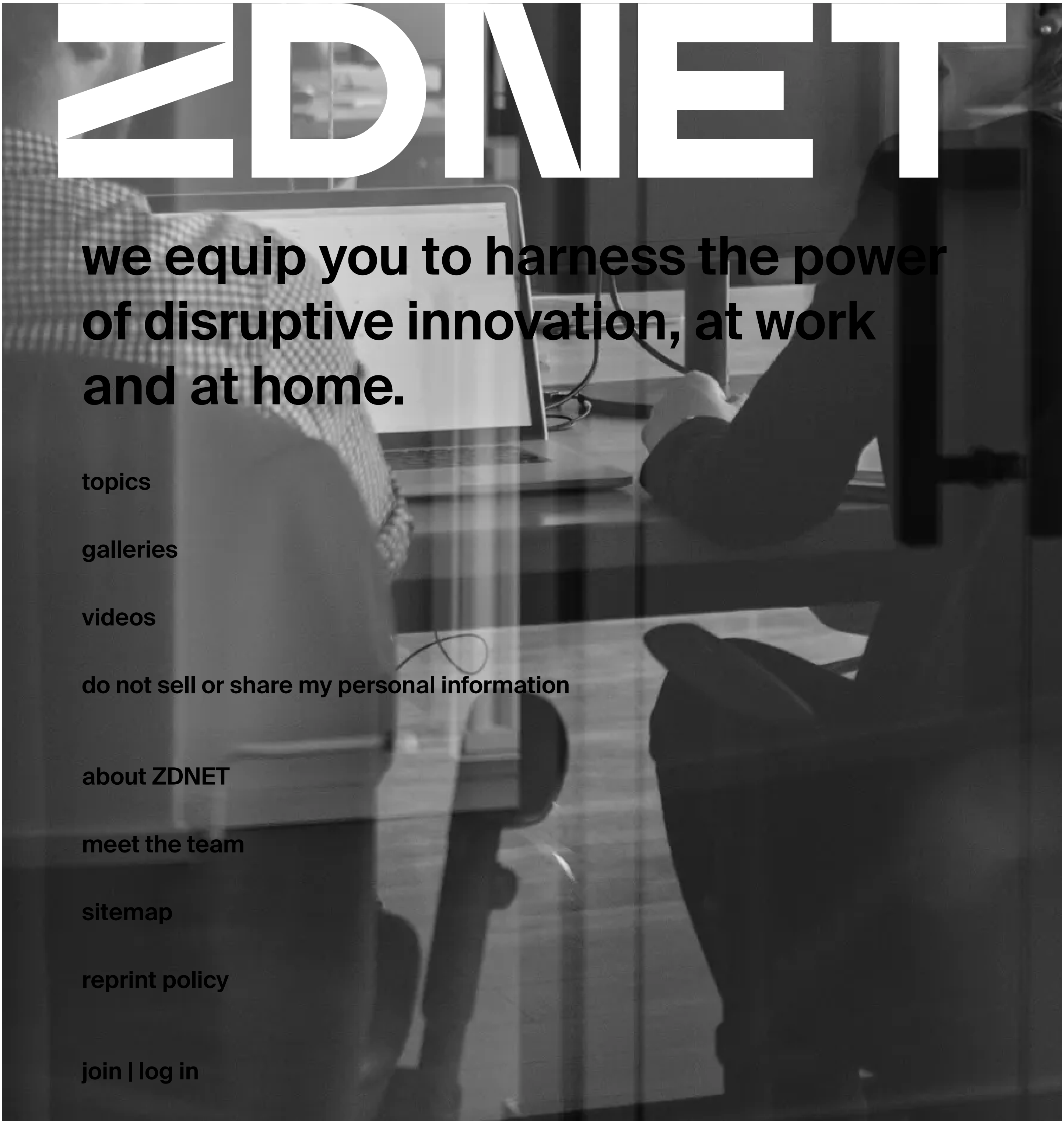
/ artificial intelligence

The impact of artificial intelligence on software development? Still unclear

Android 14's AI-generated wallpapers are super fun. Here's how to create the

Editorial standards

show comments ↓



we equip you to harness the power of disruptive innovation, at work and at home.

topics

galleries

videos

do not sell or share my personal information

about ZDNET

meet the team

sitemap

reprint policy

join | log in

newsletters

site assistance

licensing