Making GenAI more cost-efficient and accessible with open-source LLMs

Author: Yuri Sá

Open-source large language models (LLMs) have improved significantly in the past twelve months in terms of performance, developer experience, and community support. Does this mean they can offer a more cost-efficient, flexible, and secure alternative to commercial platforms? Let's explore what advantages can make open-source LLMs a viable solution for your company in 2024.

What do people want from LLMs?

Generative AI and LLMs are top-of-mind these days, dominating headlines, industry events like CES and MWC, and top software companies' earnings calls.

[Alphabet logo]	[Microsoft logo]	[Meta logo]
75	61	87
The numbers of mentions of AI during earnings calls at Alphabet, Microsoft, and Meta as of Q4 2023. Source: Medianama.com		

The great thing about this hype is that everyone wants to build something cool with Gen AI. As for the not-so-great thing, developing raw using only APIs, like OpenAI or Vertex, can get expensive fast.

Imagine a simple app that summarizes PDFs leveraging OpenAI's GPT-3.5 Turbo API priced at $0.0005 per thousand input tokens. Assuming interactions with an average context of 2000 tokens (roughly 1500 English words) and an average user making queries every 5 minutes for two hours on workdays, we're spending:

*$0.0005 * 2 (thousand tokens) * 24 (daily interactions) * 20 (workdays) = $0.48/month per user*

Scale this estimate to 100,000 monthly users, add the cost of ChatGPT output, and you'll get a $50,000–60,000 monthly bill. Now, imagine a more complex product that demands model fine-tuning, retraining, or retrieval-augmented generation (RAG). Estimates get difficult in this case, but we're looking at budgets only accessible to large corporations or unicorns.

So how can small and medium companies, garage startups, and even individuals develop LLM-powered products and functionality? That's where open-source LLMs come into play.

Open-source and free models

Navigating the landscape of open-source and free software can be daunting, especially when it comes to AI models. Take the LLM rockstar OpenAI as an example. The company initially was a non-profit committed to making its patents and research open to the public. Starting from GPT-3, OpenAI's LLMs are no longer open, and its top-of-the-line software isn't free.

I won't dare dive into the difference between licensing models, but let's use **"open-source" as a catch-all for when you take the code, modify it, and even call it yours**. This includes hundreds of LLM models, with prominent examples like Smaug-72B, Qwen-72B, Mixtral 8x7B, and more.

Let's also consider **free software which allows you to take the model code and run it wherever you want**. A prominent example is Meta's Llama, which you can use commercially, yet it's not open-source. While the practical differences here may seem minute, their compliance and legal implications can be massive.

Any great options out there?

The Hugging Face leaderboard is a great place to keep track of the latest and greatest in the open LLM space. A more specific recommendation — and a name worth following in 2024 — is Mistral, a French company founded by AI experts from Meta and Google DeepMind. Mistral's model Mixtral 8x7B outperforms GPT-3.5 in multiple benchmarks and is licensed under Apache 2.0.

Another solid option is Smaug-72B developed by Abacus AI. Launched in early 2024, Smaug-72B became the first open-source model to average an 80+ score across major LLM benchmarks.

	GPT-3.5 (prop)	Gemini Pro (prop)	Mistral - Small (open-source)	Mistral-Medium (prop)	Smaug-72B (open-source)
MMLU	70.0	71.8	70.6	75.3	**77.15**
HellaSwag	85.5	84.7	86.7	88.9	**89.27**
Arc	85.2	unknown	85.8	89.9	**76.02**
WinoGrade	81.6	unknown	81.2	**88**	85.05
GSM-8K	57.1	unknown	58.4	66.7	**78.7**
Truthful QA	unknown	unknown	unknown	unknown	**76.67**

Abstract from the Hugging Face leaderboard from February 2024. Source: Venturebeat

Currently, quite a few models score 80 or above on the Hugging Face leaderboard. All of these models have different strengths, weaknesses, and licenses — all of which you'll need to consider based on your specific goals and challenges.

Cutting operational costs with unified development platforms

When building Gen AI solutions, a considerable amount of expenses comes from running LLMs on their respective platforms through APIs. This is particularly true if you're training or extending a model — and is outright painful if your AI project isn't for production. In AI and data, nothing stings worse than incurring a huge cloud bill for an experiment gone wrong or a zero-revenue project.

A solution to this potentially costly challenge is finding a platform for running the same model across multiple platforms. Before you say anything, I know you can probably build neat VMs with GPU (or even TPU) access for your LLM needs, but this can quickly get clunky and impractical. You'll likely need a more flexible approach that supports different architectures and even local execution on machines with non-standard components (looking at you, Macs with Apple Silicon).

Llama.cpp is one of the options worth considering for working with LLMs across platforms. It is fully compatible with the Llama framework, can operate wherever LLama runs, and allows you to use extensions like the Llama index. And the best part is that it supports multiple models.

It's not just cost-savings: customization, training, and proprietary data

When it comes to open-source LLMs, access to source code can be game-changing in ways that go beyond cost-savings. In essence, open-source models transform businesses from passive users into active participants in the model's evolution.

With enough resources, your internal development team can dive into the very heart of the model, identifying potential biases or limitations and pinpointing areas ripe for customization. From tailoring models to your unique data and workflows to fine-tuning the generation process to achieve specific desired outcomes, the opportunities are limitless.

Compliance, copyright, and intellectual property

Some of the copyright and compliance concerns regarding LLMs in production focus on the ownership of the media objects produced by Generative AI. Who really owns the content generated by non-licensable software or closed-source solutions with a strictly commercial licensing model? Questions like this make AI-generated media a little scary for compliance officers.

With an open-source model, you own the tool. As a result, your compliance team has one less potential issue to worry about.

Sensitive data and local training

Local training is a crucial advantage that open-source LLMs offer for handling sensitive data. Since you have access to the model's code, you can train it on your data and within your infrastructure. This keeps sensitive information entirely under your control, mitigating privacy concerns and compliance risks often associated with cloud-based training offered by commercial LLM providers.

Closing thoughts

The future of LLMs is open and inclusive. Open-source models democratize access to Gen AI development, allowing startups, SMBs, and individuals to join the AI revolution. While cost savings are often the primary driver for open-source LLM adoption, these solutions also offer other strengths. In particular, the potential advantages in terms of compliance and local training on proprietary data make open-source LLMs attractive for big companies.

If you've found this article relevant to your AI development strategy, feel free to contact our AI Studio for practical recommendations focusing on your specific case. From choosing optimal architecture and technology to in-depth implementation strategies, we're always happy to share our expertise and insights!