



The Business Case for

LLAMA 3

in Modern Data & AI Workflows



LLAMA 3

*The Business Case for LLAMA 3 in
Modern Data & AI Workflows*

Introduction

Why LLAMA 3?

With so many open-source LLMs flooding the market, it is hard to know which one can truly move the needle for your business. The latest from Meta – LLAMA 3– has brought open source on par with the best commercial LLMs. While GPT still holds the top spot in terms of accuracy and performance, LLAMA 3 is catching up quickly. LLAMA 3-8B, for example, outperforms models ten times its size, like its predecessor LLAMA 2-70B. Once LLAMA 3-405B finishes training, it's expected to match or even surpass the latest version of GPT-4, making it the best open-source LLM available.

Why Use an Open Source LLM?

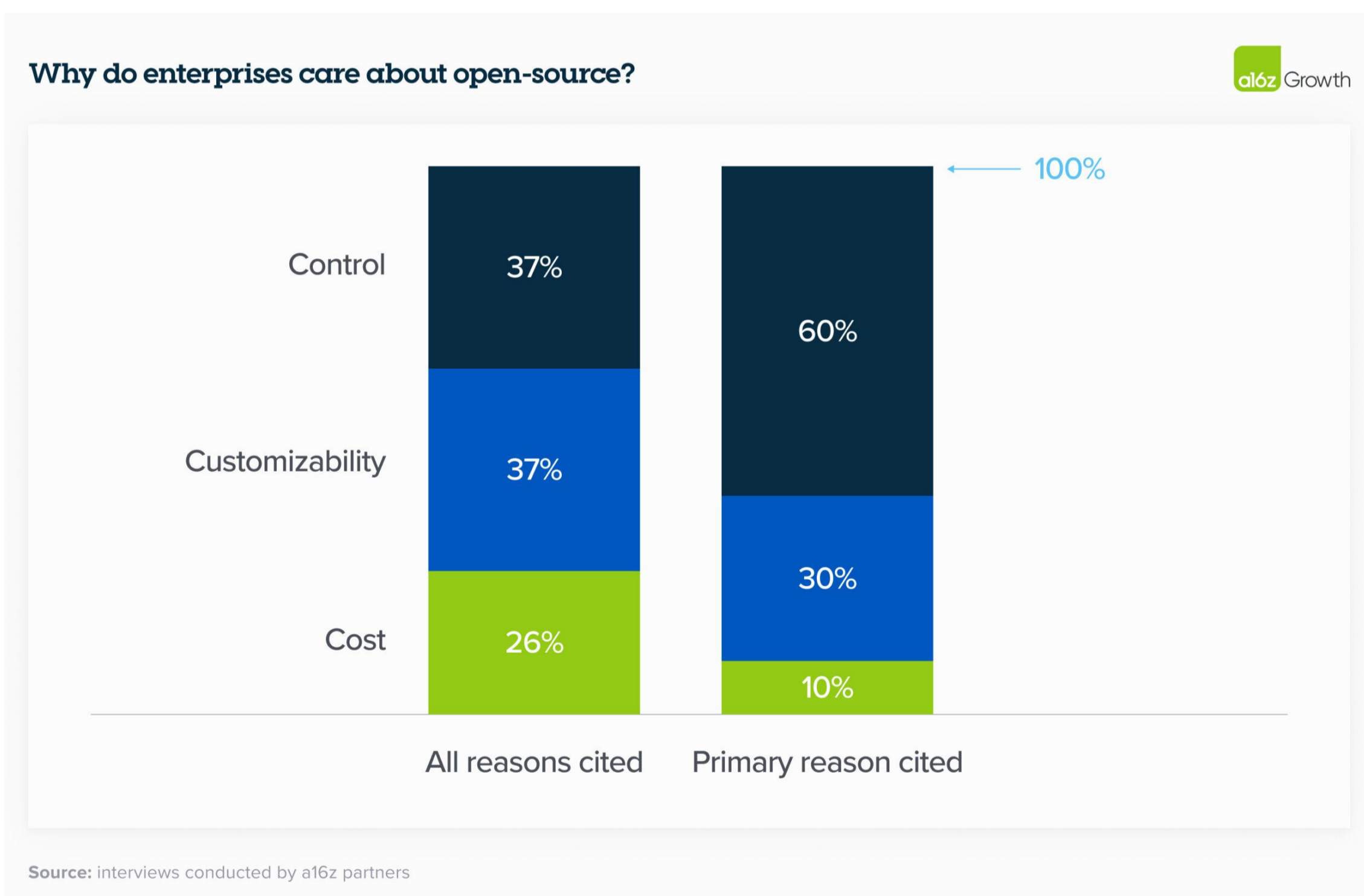
According to [Andreessen Horwitz](#), the market was dominated by closed-source models in 2023, with OpenAI holding the majority share. However, 46% of the surveyed respondents preferred open-source models moving into 2024. Interviews with AI leaders show that 60% are interested in increasing or switching to open-source usage once fine-tuned models match the performance of closed-source ones. Some are aiming for an even 50/50 split—up from the 80/20 split in favor of closed models in 2023.

Why Are Companies Making This Move?

Andreessen Horowitz's survey of Fortune 500 leaders identified three key reasons:

- 1. Control:** More control over the model and organizational data
- 2. Customizability:** Easier to fine-tune and adapt to specific use cases and needs
- 3. Cost:** Lower costs compared to other proprietary models

LLAMA 3 is an extremely competitive model in all three categories.



Source: <https://a16z.com/generative-ai-enterprise-2024/>

Control

LLAMA 3, under the [Meta Llama 3 Community License Agreement](#), allows almost all commercial uses. Key points:

- You need a license for applications with over 700M monthly active users
- You cannot use the outputs to train competing models
- No use of Meta trademarks

For most businesses, these conditions are minor. If your application exceeds 700M MAU, you can request a license from Meta.

The alternative would be MIT or Apache 2.0 licensed models. Unfortunately, there are no Apache 2.0 or MIT-licensed models within the top ten models based on Huggingface’s LMSys Chatbot Arena Leaderboard, and the only other non-proprietary model is not for commercial use (CC-BY-NC-4.0). This means that unlike other top models, which aren’t available under Apache 2.0 or MIT licenses, LLAMA 3 offers the right balance of control and performance.

Rank* (UB)	Model	Arena Elo	95% CI	Votes	Organization	License
1	GPT-4-Turbo-2024-04-09	1257	+4/-3	30562	OpenAI	Proprietary
1	GPT-4-1106-preview	1253	+3/-3	69871	OpenAI	Proprietary
2	Claude 3 Opus	1251	+3/-3	75684	Anthropic	Proprietary
2	Gemini 1.5 Pro API-0409-Preview	1248	+4/-3	33723	Google	Proprietary
3	GPT-4-0125-preview	1247	+2/-3	61924	OpenAI	Proprietary
6	Bard (Gemini Pro)	1209	+5/-5	12388	Google	Proprietary
6	Llama-3-70b-Instruct	1207	+4/-4	27298	Meta	Llama 3 Community
6	Claude 3 Sonnet	1202	+3/-3	75418	Anthropic	Proprietary
9	Command R+	1192	+3/-3	41262	Cohere	CC-BY-NC-4.0
9	GPT-4-0314	1189	+4/-3	48390	OpenAI	Proprietary

Source: <https://chat.lmsys.org/?leaderboard>

This table measures performance as the Arena Elo, or “ELO” rating. It includes close to 100 models, close to 1M votes, and is widely recognized as the “ground truth” of model quality. ELO is a measure popularized in chess where competitors (LLMs) are rated based on their relative skill levels against other competitors (LLMs). This is a good measure of LLM performance as benchmarks can easily be gamed (by training on the benchmark data). The performance of the LLMs in the LMSys leaderboard are crowdsourced, where users provide one query to several LLMs and select the best answer.



Andrej Karpathy ✓

@karpathy

I pretty much only trust two LLM evals right now: Chatbot Arena and r/LocalLlama comments section

1:44 PM · Dec 20, 2023 · **221.5K** Views

Source: <https://x.com/karpathy/status/1737544497016578453>

Control over your data is another reason to choose open-source. While data security is not unique to LLAMA 3, it is the first open-source model to rank this high in performance benchmarks. API providers like OpenAI and Anthropic have enterprise security offerings, but your data must be sent to their servers to be processed. Sending data to an API endpoint hosted outside your cluster can be risky. Data can be intercepted during transmission, third parties can gain potential unauthorized access, and external vulnerabilities can be exposed. In addition, relying on external endpoints opens your system to downtime or service disruptions.

With LLAMA 3, you don't have to send data to external servers, reducing risks like data interception or unauthorized access. Self-hosting LLAMA 3 means full control over your data, which is crucial for sensitive information.

Customizability

LLAMA 3 is the most customizable model available because of its top-tier base model performance and small parameter size, making it cost-effective to fine-tune. To illustrate this point, consider OpenBioLLM-70B, an open-source medical domain model by the team at Saama AI Labs, released just weeks after LLAMA 3 came onto the scene.

OpenBioLLM-70B is the current state-of-the-art in several biomedical tasks, beating out much larger models like Med-PaLM-2, GPT4, and Gemini-1.0. The team also trained an 8B flavor of the model, OpenBioLLM-8B, which outperforms GPT3.5 Turbo in these tests, too.

Here is a sample demonstrating the effectiveness of a fine-tuned LLAMA 3:

Question: A 35-year-old woman comes to the physician because of a 1-month history of double vision, difficulty climbing stairs, and weakness when trying to brush her hair. She reports that these symptoms are worse after she exercises and disappear after she rests for a few hours. Physical examination shows drooping of her right upper eyelid that worsens when the patient is asked to gaze at the ceiling for 2 minutes. There is diminished motor strength in the upper extremities. The remainder of the examination shows no abnormalities. Which of the most likely diagnosis?

OpenBioLLM:

Based on the symptoms and physical examination findings, the most likely diagnosis is Myasthenia Gravis (MG).

Here's why:

- **Double vision:** Diplopia is a common symptom of MG, caused by weakness of the extraocular muscles.
- **Difficulty climbing stairs:** Proximal muscle weakness, particularly in the legs, is a hallmark of MG.
- **Weakness when trying to brush her hair:** This suggests weakness in the upper extremities, which is consistent with MG.
- **Worsening with exercise and improvement with rest:** This pattern of symptoms is characteristic of MG, as the disease affects the neuromuscular junction, leading to fatigue and weakness with repeated muscle use.
- **Drooping of the right upper eyelid (ptosis):** This is a classic sign of MG, and the worsening of ptosis with sustained upward gaze (e.g., looking at the ceiling) is a diagnostic feature.
- **Diminished motor strength in the upper extremities:** This finding supports the diagnosis of MG, as the disease often affects the muscles of the arms and hands.

Myasthenia Gravis is an autoimmune disorder that affects the neuromuscular junction, leading to muscle weakness and fatigue. The diagnosis is typically made based on a combination of clinical features, electromyography (EMG), and serological tests (e.g., acetylcholine receptor antibodies).

Source: <https://chat.lmsys.org/?leaderboard>

Once fine-tuned, these models are high-performing, and thanks to techniques like LoRA and QLoRA, fine-tuning is relatively cheap. Examples of fine-tuning LLAMA 3-8B and LLAMA 3-70B for just tens or hundreds of dollars are readily available online.

Comparatively, fine-tuning with OpenAI currently requires a minimum spend of \$2-3M. Anthropic, Cohere, and similar foundational model providers could be half as expensive and still put the costs of customizability for commercial models north of \$1M. In addition, OpenAI suggests buying billions of tokens to get started.

Fine-tuning LLAMA 3 is inexpensive, and the results can lead to state-of-the-art performance. The results achieved here are impossible for most companies through providers like OpenAI but will likely become achievable using open-source LLMs, thanks to LLAMA 3.

Cost

Cost is measured as Price/Performance. Price is the cost of 1M tokens of inference based on standard pricing for commercial models and an average across inference providers for OS.

The top 7 ELO-rated models from our earlier analysis (only the most recent GPT4 model is included here) highlight LLAMA 3-70B and Gemini 1.0 Ultra as the clear price/performance leaders.

Gemini 1.0 Pro provides ten times more intelligence per dollar than its peers and twenty times more than the leaders Claude 3 and GPT4. Gemini 1.0 Pro is the loss leader within the group of highest-performing commercial models. With that in mind, LLAMA 3-70B matches the loss leader in price/performance, while being many times smaller (parameter count), and open-source.

Once again, LLAMA 3-70B is at the top of the benchmark.

The Case for a Data & AI Operating System

To recap, many businesses don't want to send their data to OpenAI or be locked into token-based pricing. LLAMA 3, as an open-source model, removes these constraints, allowing you to use a powerful LLM for various use cases. However, running LLAMA 3 within your infrastructure requires DevOps expertise, takes time, and offers no guarantee of success.

But what if you could have it all? A structured data and AI Operating System enables a true end-to-end data workflow that can be deployed on any cloud or on-prem. Using a Data and AI OS, you can easily integrate LLAMA 3 into your setup, simplifying the use of generative AI across your operations.

Benefits of Using Llama 3 on Data & AI OS

Time to Value

Without an OS, the time to value with open AI is actually shorter than the time to value with LLAMA 3. OpenAI is just an API. With LLAMA 3, you have to set up a server, deploy it, connect to it, and configure it, so more work is required before you can start getting value out of it. Add to that the fact that it's not quite as good as GPT so more prompt engineering is needed, making the time to value with LLAMA 3 longer than OpenAI, or any SaaS model.

With an OS, the time to value is much shorter because the experience is the same as using OpenAI—you just use it. So, for the many use cases where LLAMA 3 is accurate enough, the OS essentially puts it on par with an easy-to-use SaaS product like GPT in terms of time to value.

Data Security

For highly secure environments, LLAMA 3 on an OS is preferable. It ensures full control over your data, unlike using external API endpoints which pose security risks. This model also significantly reduces the compliance burden as the data does not leave your controlled environment. Running LLAMA 3 inside your Virtual Private Cloud (VPC) ensures that all data querying happens within your secure network infrastructure, minimizing the risk of unauthorized access and data breaches.

DevOps Resources

Building a custom AI infrastructure requires costly and time-consuming DevOps resources. A typical project is likely to take about 24 months of engineering effort, plus the ongoing salaries for 2-3 DevOp engineers, which can average \$300-400K. A data and AI OS with fixed fees can significantly reduce these costs and eliminate the need for extensive DevOps resources, making it a more viable option for working with LLAMA 3.

Easier Maintenance

AI evolves quickly. What's hot today might be outdated in two weeks or six months from now. An OS provides frictionless access to the latest tools, ensuring you always have the best models without the headache of upgrades and migrations. An OS allows you to keep the door open for better LLM models to be swapped in the future.

LLAMA 3 and a Data & AI OS are a Winning Combination for Modern Enterprises

LLAMA 3 excels in control, customizability, and cost. By using a Data and AI Operating System, you can achieve time to value similar to GPT, retain control over your data, avoid token-based pricing, and stay open to future model upgrades. By using a data and AI OS, you get LLAMA 3 in-house, prepare your data for fine-tuning, and deploy Llama 3 for your internal and external business applications with minimal engineering effort – thanks to tools like [Ollama](#) for hosting LLM inference, [Airbyte](#) for data ingestion, and [H2O.ai](#) for LLM finetuning – all available on a data and AI OS like Shakudo and deployed directly on your infrastructure.

With no additional DevOps or engineering work required, Shakudo brings all the tools you need to accelerate and scale your data and AI stack. With a data and AI OS, you can reap the rewards of groundbreaking AI technology like LLAMA 3 in weeks, not months.



ABOUT SHAKUDO

Shakudo creates compatibility across the best-of-breed data tools for a more reliable, performant, and cost effective data and AI operating system. As an operating layer on top of your cloud Shakudo allows you to pick the best-of-breed data tools for your needs, while providing a platform with fully automated DevOps experience. This combines the best of both worlds in data stack practices so you can focus on delivering business value with data.

Shakudo is the most **easy, secure, cost-effective, scalable** way to bring the most advanced data and AI tools to your data. Find out more at **shakudo.io**.