The Executive's Guide to LLMs: Open-Source vs Proprietary

run:
ai

The Executive's Guide to LLMs: Open-Source vs Proprietary

As artificial intelligence (AI) continues to transform industries, choosing between open-source and proprietary AI models is a critical decision for businesses looking to leverage AI's full potential. Both options come with distinct advantages—ranging from control and flexibility to scalability and support—and each plays a unique role in shaping the future of AI-driven innovation.

In today's rapidly evolving AI landscape, the question "Which model should my team choose?" is common, but often misses the mark. The real challenge lies in understanding the broader context and implications of AI model selection. This guide aims to equip decision-makers with essential questions to ask when considering AI models, key factors to be aware of in the selection process and strategies for aligning AI capabilities with organizational goals.

In this guide, you will learn:

> How to align AI model selection with your business strategy

> Why understanding AI model licensing is crucial for cost management

> How to assess AI models for scalability and future-proofing

> How to navigate the hidden costs of AI model deployment

The Evolution of AI Models: A Brief Overview

AI models are now at the core of cutting-edge technologies—from natural language processing (NLP) and computer vision to generative AI tools that can create text, images, and even code. The models in this ecosystem can be broadly divided into two categories: open-source and proprietary models.

Open-source models are freely available for download, modification, and deployment, offering businesses complete control over customization and usage within the bounds of licenses.
Proprietary models are developed and controlled by companies, typically offered through APIs or licensed platforms, often with additional enterprise-level support and pre-built integrations.
For today's business leaders, the challenge is understanding the trade-offs between the customizability and control of open-source models and the ease-of-deployment and enterprise support provided by proprietary solutions. Let's explore the key players in each space.

Key Open-Source AI Model Providers

Meta (LLaMA Series)

Meta's LLaMA models are powerful large language models (LLMs) designed for efficiency and scalability. The latest addition, the Llama 3.2 series, introduces advanced capabilities with a focus on edge and mobile device use cases. In general, Llama model series come in various sizes, from 1B to 405B parameters, making them suitable for businesses that need to deploy AI across a range of environments, including low-resource settings. LLaMA's flexibility makes it popular among researchers and enterprises that want to fine-tune models for specialized tasks.

Mistral

Mistral is a rising European AI company focused on delivering high-performance generative models. Although relatively new, their focus on efficiency and optimization makes their models appealing for organizations looking for scalable, open-source alternatives to proprietary LLMs.

Side note: While most of the Mistral models are licensed under Apache 2.0, some (e.g. Codestral) are under a different license. These models can be used for non-commercial purposes with no strings attached. A commercial license must be purchased for commercial use cases requiring self-deployment. For more information about Mistral Research License and Mistral Commercial License, please refer to their pricing page.

Cohere for AI

Cohere for AI is the open research arm of Cohere, focusing on multilingual language models and natural language understanding. While Cohere's commercial models are proprietary, their open-source contributions are geared toward language model training and embedding generation.

Stability AI (Stable Diffusion)

Stability AI has democratized text-to-image generation through its Stable Diffusion models, which are widely used in industries such as media, design, and entertainment. Stability AI's commitment to open-source innovation has made it a go-to for businesses looking to create high-quality generative models.

Note, there is a growing number of additional open-source providers, including but not limited to Google, Microsoft, Nvidia, BigScience, LAION, EleutherAI, Qwen and BAAI.

Key Proprietary AI Model Providers

OpenAI (GPT Models)

OpenAI's GPT models (including GPT-4 and o1) are among the most advanced and widely used for tasks like chatbots, code generation, and image synthesis. However, these models are closed-source, available primarily through APIs via OpenAI's platform or Microsoft Azure. This makes them a convenient option for rapid deployment, but customization and data privacy may be limited.

Google (Gemini Models)

Google's Gemini models, the successor to PaLM, power AI-driven NLP and conversational tools integrated within Google's enterprise ecosystem. Available exclusively through Google Cloud, these models provide high scalability and are designed to meet the needs of enterprises looking for plug-and-play AI solutions.

Other Proprietary Providers:

- AWS Bedrock: AWS Bedrock offers access to several proprietary models such as Titan (AWS), Claude (Anthropic), and Jurassic-2 (AI21 Labs). While these models are generally API-driven, SageMaker Edge Manager provides options for on-premises or edge deployments, catering to enterprises requiring hybrid infrastructure setups.
- Cohere Commercial Models: Available via API with enterprise plans for private deployment.
- Palantir Foundry: Custom AI models that can run in hybrid cloud and private data center environments.

Open-Source vs. Proprietary Models: Key Considerations for Enterprises

Proprietary Models: Quick Start, Long-Term Trade-offs

Proprietary models, accessible through APIs, offer enterprise-level support and a quicker time to market. They are ideal for businesses looking for a rapid AI deployment strategy without needing extensive in-house expertise.

However, as companies scale, the limitations of proprietary models become apparent:

- **Data Privacy**: Customization often requires sending sensitive data to third-party vendors, which poses potential security risks.

- **Costs**: While pay-per-usage pricing can be appealing initially, the costs can quickly balloon as AI adoption grows.

- **Vendor Dependency**: Businesses become dependent on the reliability and uptime of external infrastructure, which can pose operational risks if a vendor experiences outages.

Open-Source Models: Control, Flexibility, and Cost Efficiency

On the other hand, open-source models allow businesses to fully control their AI infrastructure. These models can be fine-tuned using proprietary data, giving companies the ability to tailor their AI performance to specific needs without sharing sensitive data externally.

- **Customization**: Fine-tuning open-source models allows for deep customization, making them a better fit for industry-specific applications.

- **Speed**: Since the models are customizable, one can also optimize the model depending on the latency and hardware requirements.

- **Cost Efficiency**: By deploying models on-premises, companies can avoid the escalating inference costs associated with usage-based proprietary models.

- **Flexibility**: Open-source models can be deployed across any infrastructure—whether on-premises, cloud, or hybrid environments—allowing for full scalability at the pace of the business.

Ultimately, the power of open-source models isn't just in their accessibility or cost-effectiveness—it's in their adaptability. By investing in fine-tuning with proprietary data that aligns with strategic goals, organizations can innovate faster and cheaper in the long term.

As noted, open-source models offer many benefits. However they also come with some challenges:

Orchestration at Scale

Open-source AI models are dynamic, often requiring sophisticated orchestration as your AI footprint grows. Managing multiple models across teams and hardware infrastructures demands policies and optimized resource allocation in clusters.

Upfront Investment

While open-source models are often free, businesses may face upfront costs in setting up the necessary infrastructure and expertise.

Licensing Considerations

Open-source AI models aren't always free for commercial use. Some, like Mistral AI's, have custom licenses that may require payments for certain deployments. Always check the licensing terms before use.

To overcome these challenges and simplify use of open-source models, many tools and solutions have been developed, both open source and proprietary.

Now that we have a better understanding of advantages and disadvantages of both worlds, let's go through the key questions and considerations that actually matter the most when choosing between them.

Reframing the Question, "Which model should I choose?"

The right model isn't something you can pick off a shelf. Instead of searching for the perfect fit, focus on experimenting with different models on your own data. Benchmark the results to find the best match for your use case.

How should you benchmark the models? Public leaderboards, like Hugging Face's, are a great starting point for evaluating models across tasks. However, it's important to create internal benchmarks based on your unique data to determine which model performs best for your specific needs.

As you evaluate the right fit for your enterprise, keep in mind:

! The "Bigger is Better" Myth: The Role of Fine-Tuning

A common misconception is that larger AI models always deliver better performance. However, models like LLaMA-3 405B are foundational models—designed to handle a wide range of tasks, but without specialization. In many cases, smaller models, such as LLaMA-3 8B, outperform larger ones when fine-tuned for specific use cases, such as legal document processing or financial report interpretation. The key lies not in size, but in how well the model is tailored to your unique business needs.

! Customization: The Key to AI Success

Following up on the previous point; Fine-tuning allows businesses to get insights from their data in a meaningful way by training the foundational models on domain-specific data, transforming them from generalists into specialists that deliver superior performance in targeted areas.

For example, a foundational AI model can understand natural language, but it won't be fluent in legal terminology or financial analysis unless you tailor it with your own proprietary data. This customization is where AI delivers a real competitive advantage—allowing you to automate tasks, improve decision-making, and drive innovation in ways that matter most to your organization.

Side note: While some models are pre-trained on specialized data (e.g., healthcare or finance), fine tuning with proprietary data may still be required.

! Watch Out for Licensing

As mentioned previously with the Mistral AI case, remember that not all of the models on HuggingFace or any other platform are truly free. Some models have custom licenses that may require a commercial agreement. Make sure to review these terms carefully before integrating them into your workflows to avoid unexpected costs down the line.

Looking Ahead

For organizations looking for greater control, flexibility, and cost efficiency, open-source models offer a clear path forward. Unlike proprietary models, which often come with high costs when scaled, limited customization, and dependency on third-party infrastructure, open-source solutions put the power back in your hands. With the ability to fine-tune and deploy models on your own terms—whether on-premises, in the cloud, or a hybrid setup—you gain the freedom to innovate at your own pace.

While there may not be a clear-cut answer to the question, "Which model should I choose?", we hope that this guide will help you to get a better understanding in which direction you should look when making these decisions.

Ready to maximize the potential of GenAI while maintaining control, security, and scalability. Here's how Run:ai can help:

› **Accelerate Innovation with Open-Source Models:** Deploy models like LLaMA and Mistral on your own infrastructure, maintaining full control over data and compliance.

› **Optimize Performance and Costs**: Minimize cold start delays and resource over-provisioning, ensuring faster model readiness and improved cost efficiency.

› **Scale Your AI Lifecycle Seamlessly**: Manage resources dynamically across hybrid environments, from development to inference, maximizing GPU utilization.

Take the next step in your
AI journey with Run:ai.

www.run.ai

run:
ai