What is a llama? Why is it important?

Llama is a family of large language models (LLMs) and large multimodal models (LMMs) from Meta. The latest version is Llama 4.

Llama is a family of large language models (LLMs) and large multimodal models (LMMs) from Meta. The latest version is Llama 4. Essentially, it's Facebook's response to OpenAI and Google Gemini – but with one key difference: all Llama models are freely available for almost anyone to use for research and commercial purposes.

This is a crucial aspect, and it has made Llama models incredibly popular with AI developers. Let's explore what Meta's Llama models offer!

What is a llama?

Llama is a family of LLM models (and those with image processing capabilities, or LMMs) similar to OpenAI's GPT and Google's Gemini. Currently, the version numbers are a bit mixed. Meta is at Llama 4 for some models, and Llama 3.3, 3.2, and 3.1 for others. As more Llama 4 models are released, it's possible that other Llama 3 models will be phased out – but they are currently still available and supported.

At the time of writing, the models available for download from Meta are:

1. Llama 3.1 8B
2. Llama 3.1 405B
3. Llama 3.2 1B
4. Llama 3.2 3B
5. Llama 3.2 11B-Vision
6. Llama 3.2 90B-Vision
7. Llama 3.3 70B
8. Llama 4 Scout
9. Llama 4 Maverick

Additionally, there are two unreleased Llama 4 models:

1. Llama 4 Behemoth
2. Llama 4 Reasoning

In general, all Llama models operate on the same basic principle. They use variations of the transformer architecture and are developed using pre-training and refinement. The biggest difference is that Llama 4 models are inherently multimodal and use a Mixture-of-Experts architecture.

When you input a text prompt or provide the model with input data in some other way, it attempts to predict the most likely next piece of text using its artificial neural network—a cascading algorithm with billions of variables (called "parameters") modeled after the human brain. A similar process occurs with images for the models that support them.

Different Llama 3 models offer different trade-offs between price and performance. For example, smaller models like Llama 3.1 8B and Llama 3.2 3B are designed to run on edge devices like smartphones and computers, or to operate extremely quickly and cost-effectively on more powerful hardware. The largest model, Llama 3.1 405B, offers the highest performance in most cases, but it requires the most resources to run. The Vision models are for multimodal use cases, and the Llama 3.3 70B offers an excellent balance between performance and cost.

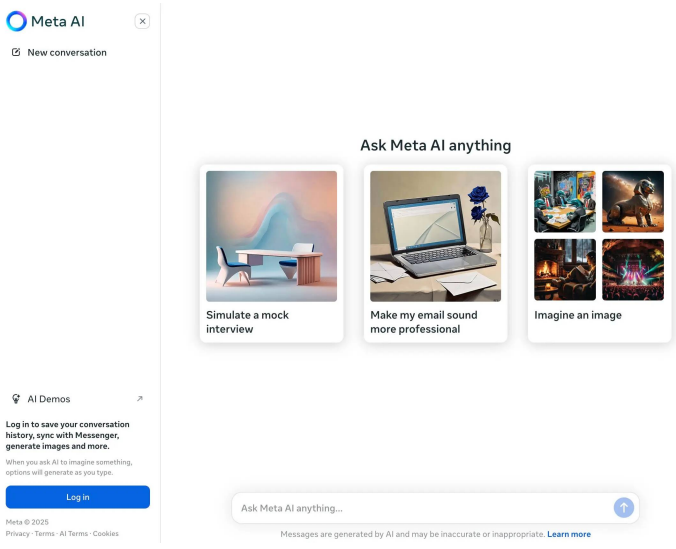
The two Llama 4 models—Llama 4 Scout and Llama 4 Maverick—use a slightly different parametric approach called Mixture-of-Experts (MoE). Llama 4 Scout has a total of 109 billion parameters but only uses 17 billion at a time. Llama 4 Maverick has a total of 400 billion parameters, but also only uses a maximum of 17 billion. This approach allows for more powerful and efficient AI models, although they are more complex to develop.

In addition to Scout and Maverick, Meta has also announced Llama 4 Behemoth. It also uses the MoE architecture and has a total of 2 trillion parameters with 288 billion active parameters. It is still in training.

It's worth noting that the announcement for Llama 4 doesn't mention any reasoning models. There's an introduction page, so it'll be coming soon, but for now, the llama population is limited to models without reasoning abilities.

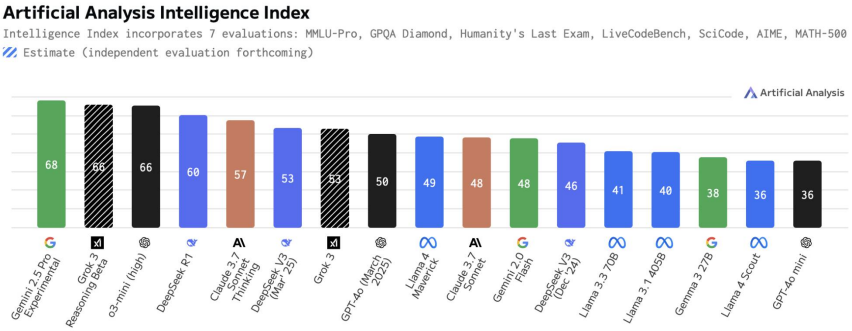
Meta AI: How to Try Llama

Meta AI, the AI ??assistant integrated into Facebook, Messenger, Instagram, and WhatsApp, now uses Llama 4. The best place to experience it is through the dedicated web application.



Comparing Llama to GPT, Gemini, and other AI models: How do they work?

Llama 4 Maverick and Scout are stable open-source models, although their performance isn't the best. Specifically, the lack of inference models (to date) prevents them from topping most performance tests.



Llama 4 Maverick competes with DeepSeek V3, Grok 3, GPT-4o , Claude Sonnet 3.7, and Gemini 2.0 Flash. As you can see in the chart above from Artificial Analysis, this is a pretty good non-inferential model, although its main advantage is being the highest-performing open-source multimodal model – and the highest-performing non-Chinese open-source language model.

Maverick's MoE structure also helps save on operating costs, especially when compared to proprietary models like GPT-4o. A test version is currently second in the chatbot field, so it's certainly showing a lot of promise. It has a context window of one million tokens, which is quite good, but other models also achieve similar levels.

The Llama 4 Scout competes well with the GPT-4o mini, but it's interesting in two respects. First, it's designed to run on a single H100 GPU. While this is still a server-class GPU, larger models typically run on a cluster of multiple GPUs rather than a single dedicated GPU. Second, it has a 10 million token context window, which is truly best in its segment. However, it's worth noting that no vendors currently support this feature.

Although Meta has released some provisional performance scores for Behemoth – clearly outperforming GPT-4.5 in a few tests – things move so fast in the AI ??field that it's not worth focusing too much on them until the official product launches. Similarly, any Llama 4 inference model would be a major step forward.

The Llama 4 is clearly the future of the Llama line, but the Llama 3 models are still good options. They're no longer considered to offer the most advanced performance, but they can be affordable and efficient.

Why are llama important?

Most of the large-scale modeling languages ??(LLMs) you've heard of—OpenAI's o1 and GPT-4o, Google's Gemini, Anthropic's Claude—are proprietary and closed-source. Researchers and businesses can use the official APIs to access them and even fine-tune model versions to provide appropriate responses, but they can't truly interfere with or understand what's going on inside.

However, with Llama, you can download the model instantly, and as long as you have sufficient technical skills, you can run it on a cloud server or simply learn its source code. You can run Llama 3 models on some computers, although Llama 4 Scout and Maverick are too large for home use.

And what's even more useful is that you can also run it on Microsoft Azure, Google Cloud, Amazon Web Services, and other cloud infrastructures to run LLM-based applications, or train it on your own data to generate the type of text you need. Just be sure to check Meta's guidelines on using Llama responsibly – the license isn't quite as lenient as a traditional open-source license.

However, by continuing to be open with Llama, Meta is making it significantly easier for other companies to develop AI-based applications over which they have more control – provided they adhere to the accepted usage policy. Worryingly, users in the EU are currently banned from using Llama 4, but we'll see if that changes once it's rolled out. The only other major limitation of the license is that companies with more than 700 million monthly users must apply for special permission to use Llama, so companies like Apple, Google, and Amazon have to develop their own LLMs.

In a letter accompanying the release of Llama 3.1, CEO Mark Zuckerberg was extremely transparent about Meta's plans to keep Llama open:

"I believe that open source is essential for a positive future of AI. AI has more potential than any other modern technology to increase human productivity, creativity, and quality of life – and to drive economic growth while unlocking advances in medical and scientific research. Open source will ensure that more people around the world can access the benefits and opportunities of AI, that power is not concentrated in the hands of a few companies, and that the technology can be deployed more evenly and securely across society."

And indeed, that's quite interesting – provided the EU issue is resolved satisfactorily. Certainly, Meta will benefit by taking a somewhat leading role in one of the most important AI models. But independent developers, companies unwilling to be tied to a closed system, and everyone else interested in AI will also benefit. Many of the major advancements in computing over the past 70 years have been built on open research and experimentation, and now AI seems to be one of them. While Google , OpenAI, and Anthropic will always be players in this field, they won't be able to build the commercial barriers or customer binding that Google has achieved in search and advertising.

By bringing Llama to market, there will likely always be a reliable alternative to closed-source AI.