# Meta Llama 4: Features, how to access it, how it works, etc.

Meta has announced the Llama 4 model lineup, which includes two already released models – the Llama 4 Scout and the Llama 4 Maverick – and a third model still in training: the Llama 4 Behemoth.

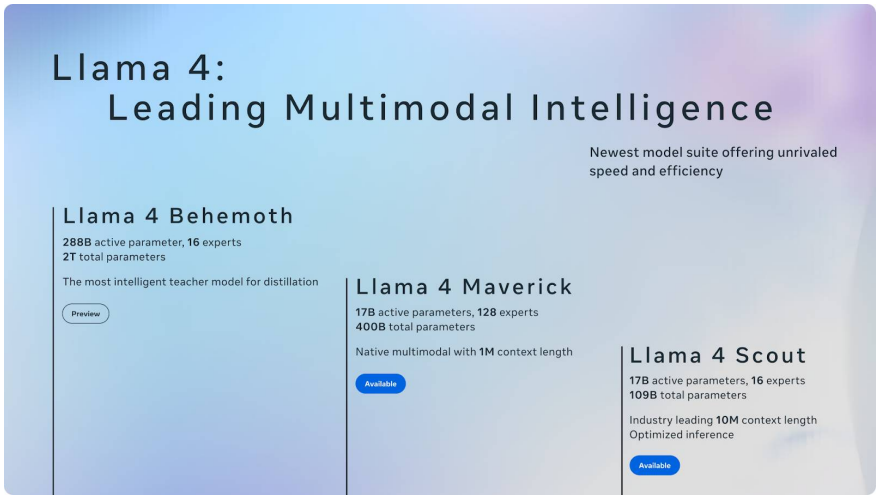
Meta has announced the Llama 4 model lineup, which includes two already released models – the Llama 4 Scout and the Llama 4 Maverick – and a third model still in training: the Llama 4 Behemoth.

The Scout and Maverick versions are now available, released publicly under Meta's standard weighted open license – with one notable caveat: If your service exceeds 700 million monthly active users, you will need to apply for a separate license from Meta, and Meta may or may not grant the license at their discretion.

Llama Scout supports a 10 million token context window, the largest among publicly released models. Llama Maverick is a versatile model and targets GPT-40, Gemini 2.0 Flash, and DeepSeek-V3. Llama Behemoth, still under training, serves as a high-capacity teaching model.

## What is Llama 4?

Llama 4 is Meta's new large-scale language modeling suite. This version includes two existing models – Llama 4 Scout and Llama 4 Maverick – and a third model, Llama 4 Behemoth, which is currently under development.

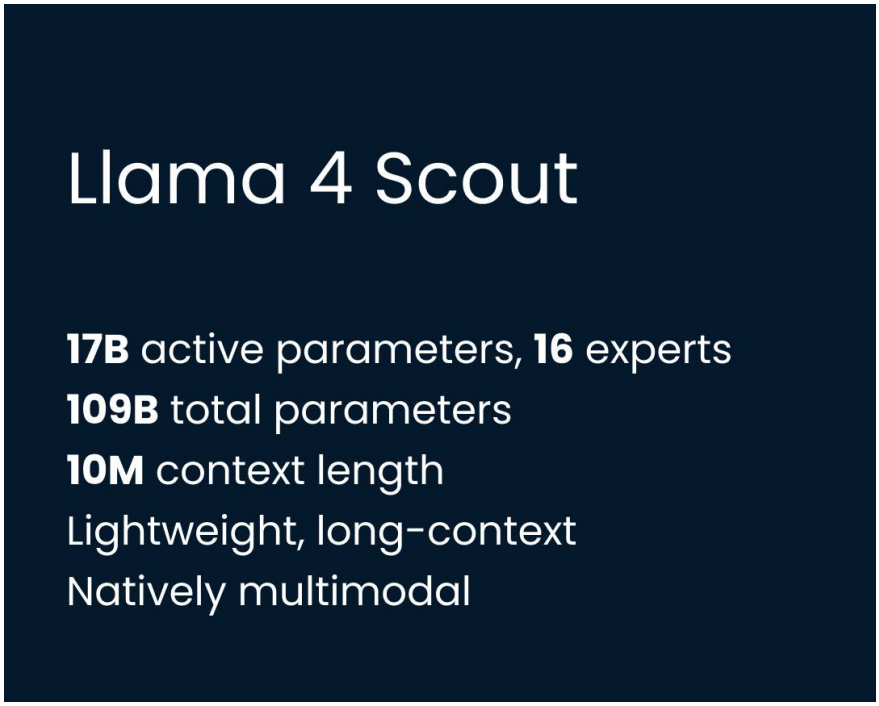


Llama 4 introduces significant improvements. In particular, it incorporates a Mixture-of-Experts (MoE) architecture, aiming to improve efficiency and performance by only activating the components necessary for specific tasks. This design represents a shift towards more scalable and specialized AI models.

Let's explore each model in more detail.

## Llama Scout

Llama 4 Scout is the lighter-weighted model in the new toolkit, but arguably the most interesting. It runs on a single H100 GPU and supports a 10 million-token context window. This makes Scout the most context-intensive open-weighted model released to date and potentially the most useful for tasks such as summarizing multiple documents, inferring long code, and analyzing operations.



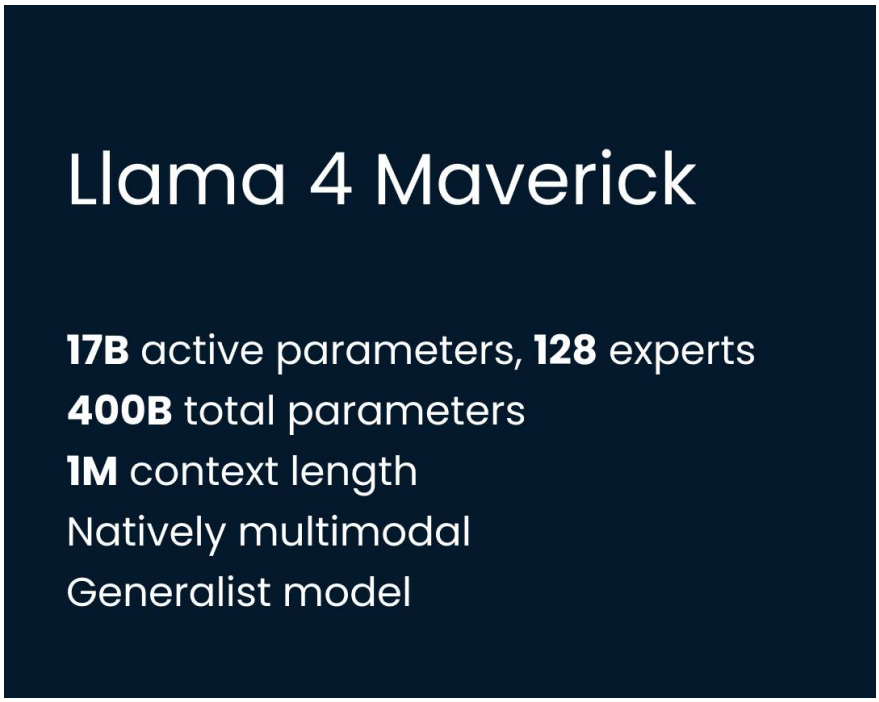

Scout has 17 billion operational parameters, organized across 16 experts, for a total of 109 billion parameters. It's pre- and post-trained with a 256K context window, but Meta claims it has far better generalization capabilities than that number suggests (this claim still needs verification). In practice, this opens the door to workflows involving entire codebases, session histories, or legal documents—all processed in a single data pass.

Architecturally, Scout is built using Meta's MoE framework, where only a subset of parameters are enabled per token—in contrast to dense models like GPT-4o , where all parameters are enabled. This means it is both computationally efficient and highly scalable.

## Llama Maverick

Llama 4 Maverick is the most versatile model in the product line – a full-scale multimodal model built to work efficiently across the domains of conversation, inference, visual comprehension, and code. While Scout pushes

the limits of context length, Maverick focuses on balanced, high-quality output across tasks. It's Meta's answer to GPT-40, DeepSeek-V3, and Gemini 2.0 Flash.

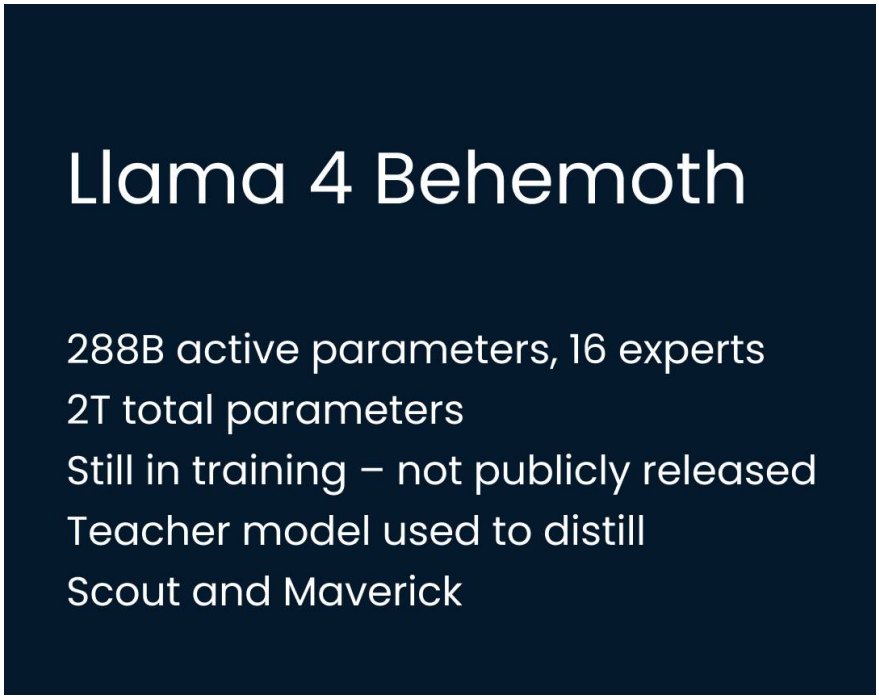



Maverick has the same 17 billion operational parameters as Scout, but with a larger MoE configuration: 128 experts and a total of 400 billion parameters. Like Scout, it uses the MoE architecture, only enabling a portion of the model per token – reducing inference costs while scaling capacity. The model runs on a single H100 DGX server, but can also be deployed with distributed inference for larger-scale applications.

## Llama Behemoth

Llama 4 Behemoth is Meta's most powerful and largest model to date – but it hasn't been released yet. Because it's still under training, Behemoth isn't a reasoning model in the same sense as DeepSeek-R1 or OpenAI's o3, which are built and optimized for multi-step thought process tasks.

Based on what we know so far, it also doesn't appear to be designed as a product for direct use. Instead, it functions as a teaching model, used to refine and shape both Scout and Maverick. Once released, it may allow others to refine their own models as well.




Behemoth has 288 billion operational parameters, organized across 16 experts, with a total parameter count of nearly 2 trillion. Meta has built an entirely new training infrastructure to support Behemoth at this scale. They've introduced asynchronous reinforcement learning, curriculum sampling based on prompt difficulty, and a new slicing function that dynamically balances soft and hard objectives.

# Llama 4 Benchmark

Meta has released internal benchmark results for each Llama 4 model, comparing them to both previous Llama variants and several competing open-weighted and pioneering models.

This section will guide you through the benchmark highlights of Scout, Maverick, and Behemoth, using Meta's own metrics. These scores provide a useful first glimpse into each model's performance across different tasks and where they stand in the current landscape. Let's start with Scout.

## Benchmark Llama Scout

Llama 4 Scout performed well on many reasoning, programming, and multimodal tests—especially considering the smaller number of operational parameters and the use of only one GPU .

Llama 4 Scout instruction-tuned benchmarks

| Category / Benchmark | Llama 4 Scout | Llama 3.3 70B | Llama 3.1 405B | Gemma 3 27B | Mistral 3.1 24B | Gemini 2.0 Flash-Lite |
|---|---|---|---|---|---|---|
| Image Reasoning — MMMU | 69.4 | No multimodal support | No multimodal support | 64.9 | 62.8 | 68.0 |
| MathVista | 70.7 | | | 67.6 | 68.9 | 57.6 |
| Image Understanding — ChartQA | 88.8 | | | 76.3 | 86.2 | 73.0 |
| DocVQA (test) | 94.4 | | | 90.4 | 94.1 | 91.2 |
| Coding — LiveCodeBench (10/01/2024-02/01/2025) | 32.8 | 33.3 | 27.7 | 29.7 | — | 28.9 |
| Reasoning & Knowledge — MMLU Pro | 74.3 | 68.9 | 73.4 | 67.5 | 66.8 | 71.6 |
| GPQA Diamond | 57.2 | 50.5 | 49.0 | 42.4 | 46.0 | 51.5 |
| Long Context — MTOB (half book) eng → kgv/kgv → eng | 42.2/36.6 | Context window is 128K | Context window is 128K | Context window is 128K | Context window is 128K | 42.3/35.1[3] |
| MTOB (full book) eng → kgv/kgv → eng | 39.7/36.3 | | | | | 35.1/30.0[3] |

1. For Llama model results, we report 0 shot evaluation with temperature = 0 and no majority voting or parallel test time compute. For high-variance benchmarks (GPQA Diamond, LiveCodeBench), we average over multiple generations to reduce uncertainty.
2. For non-Llama models, we source the highest available self-reported eval results unless otherwise specified. We only include evals from models that have reproducible evals (via API or open weights), and we only include non-thinking models.
3. Specialized long context evals are not traditionally reported for generalist models, so we share internal runs to showcase Llama's frontier performance.

In terms of image comprehension, Scout outperformed its competitors: scoring 88.8 on ChartQA and 94.4 on DocVQA (test), surpassing Gemini 2.0 Flash-Lite (73.0 and 91.2 respectively) and matching or slightly outperforming Mistral 3.1 and Gemma 3 27B.

In visual reasoning tests such as MMMU (69.4) and MathVista (70.7), it also leads the group of open-weighted modules, surpassing Gemma 3 (64.9, 67.6), Mistral 3.1 (62.8, 68.9) and Gemini Flash-Lite (68.0, 57.6).

In programming, Scout scored 32.8 on LiveCodeBench, outperforming Gemini Flash-Lite (28.9) and Gemma 3 27B (29.7), although slightly lower than Llama 3.3 with 33.3 points. It's not a preferred programming model, but its capabilities are still highly regarded.

In terms of knowledge and reasoning ability, Scout scored 74.3 on MMLU Pro and 57.2 on GPQA Diamond, outperforming all other open-weighted models on both tests. These tests prioritize long-step multi-step reasoning, so Scout's strong performance here is noteworthy, especially at this scale.

## Benchmark Llama Maverick

Maverick is the most comprehensive model in the Llama 4 product line – and benchmark results reflect that. While it doesn't aim for extreme context lengths like Scout or rudimentary scales like Behemoth, it performs consistently across all key categories: multimodal inference, coding, language comprehension, and the ability to remember long contexts.

Llama 4 Maverick instruction-tuned benchmarks

| Category / Benchmark | Llama 4 Maverick | Gemini 2.0 Flash | DeepSeek v3.1 | GPT-4o |
|---|---|---|---|---|
| Inference Cost — Cost per 1M input & output tokens (3:1 blended) | $0.19–$0.49[5] | **$0.17** | $0.48 | $4.38 |
| Image Reasoning — MMMU | **73.4** | 71.7 | No multimodal support | 69.1 |
| MathVista | **73.7** | 73.1 | | 63.8 |
| Image Understanding — ChartQA | **90.0** | 88.3 | | 85.7 |
| DocVQA (test) | **94.4** | — | | 92.8 |
| Coding — LiveCodeBench (10/01/2024-02/01/2025) | 43.4 | 34.5 | **45.8/49.2**[3] | 32.3[3] |
| Reasoning & Knowledge — MMLU Pro | 80.5 | 77.6 | **81.2** | — |
| GPQA Diamond | **69.8** | 60.1 | 68.4 | 53.6 |
| Multilingual — Multilingual MMLU | **84.6** | — | — | 81.5 |
| Long Context — MTOB (half book) eng → kgv/kgv → eng | **54.0/46.4** | 48.4/39.8[4] | Context window is 128K | Context window is 128K |
| MTOB (full book) eng → kgv/kgv → eng | **50.8/46.7** | 45.5/39.6[4] | | |

1. For Llama model results, we report 0 shot evaluation with temperature = 0 and no majority voting or parallel test time compute. For high-variance benchmarks (GPQA Diamond, LiveCodeBench), we average over multiple generations to reduce uncertainty.
2. For non-Llama models, we source the highest available self-reported eval results, unless otherwise specified. We only include evals from models that have reproducible evals (via API or open weights), and we only include non-thinking models. Cost estimates are sourced from Artificial Analysis for non-Llama models.
3. DeepSeek v3.1's date range is unknown (49.2), so we provide our internal result (45.8) on the defined date range. Results for GPT-4o are sourced from the LCB leaderboard.
4. Specialized long context evals are not traditionally reported for generalist models, so we share internal runs to showcase Llama's frontier performance.
5. $0.19/Mtok (3:1 blended) is our cost estimate for Llama 4 Maverick assuming distributed inference. On a single host, we project the model can be served at $0.30-$0.49/Mtok (3:1 blended).

In visual reasoning, Maverick scored 73.4 on MMMU and 73.7 on MathVista, outperforming Gemini 2.0 Flash (71.7 and 73.1) and GPT-4o (69.1 and 63.8). On ChartQA (visual comprehension), Maverick scored 90.0, slightly higher than Gemini (88.3) and significantly higher than GPT-4o (85.7). On DocVQA, Maverick scored 94.4, on par with Scout and outperforming GPT-4o (92.8).

In terms of programming, Maverick scored 43.4 points on LiveCodeBench, higher than GPT-4o (32.3), Gemini Flash (34.5) and nearly equal to DeepSeek v3.1 (45.8).

In terms of reasoning and knowledge, Maverick scored 80.5 on MMLU Pro and 69.8 on GPQA Diamond, again outperforming Gemini Flash (77.6 and 60.1) and GPT-4o (no MMLU Pro results, 53.6 on GPQA). DeepSeek v3.1 led by 0.7 points on MMLU Pro.

## Benchmark Llama Behemoth

Behemoth hasn't been released yet, but its benchmark numbers are already impressive.

Llama 4 Behemoth instruction-tuned benchmarks

| Category / Benchmark | Llama 4 Behemoth | Claude Sonnet 3.7 | Gemini 2.0 Pro | GPT-4.5 |
|---|---|---|---|---|
| Coding — LiveCodeBench (10/01/2024-02/01/2025) | 49.4 | — | 36.0[3] | — |
| Reasoning & Knowledge — MATH-500 | 95.0 | 82.2 | 91.8 | — |
| MMLU Pro | 82.2 | — | 79.1 | — |
| GPQA Diamond | 73.7 | 68.0 | 64.7 | 71.4 |
| Multilingual — Multilingual MMLU (OpenAI) | 85.8 | 83.2 | — | 85.1 |
| Image Reasoning — MMMU | 76.1 | 71.8 | 72.7 | 74.4 |

1. Llama model results represent our current best internal runs.
2. For non-Llama models, we source the highest available self-reported eval results, unless otherwise specified. We only include evals from models that have reproducible evals (via API or open weights) and we only include non-thinking models.
3. Results are sourced from the LCB leaderboard.

On in-depth STEM tests, Behemoth performed very well. It scored 95.0 on MATH-500 – higher than Gemini 2.0 Pro (91.8) and significantly higher than Claude Sonnet 3.7 (82.2). On MMLU Pro, Behemoth scored 82.2, while Gemini Pro scored 79.1 (Claude has no reported score). And on GPQA Diamond, another test that highly values ??the depth and accuracy of factual information, Behemoth scored 73.7, outperforming Claude (68.0), Gemini (64.7) and GPT-4.5 (71.4).

In terms of multilingual comprehension, Behemoth scored 85.8 on the Multilingual MMLU, slightly higher than Claude Sonnet (83.2) and GPT-4.5 (85.1). These scores are crucial for global developers working outside of English, and Behemoth currently leads in this category.

In terms of image inference capabilities, Behemoth scored 76.1 on MMMU, outperforming Gemini (71.8), Claude (72.7), and GPT-4.5 (74.4). While this wasn't its primary focus, it still demonstrated competitiveness with leading multimodal models.

In terms of code generation capabilities, Behemoth scored 49.4 points on LiveCodeBench. This is significantly higher than Gemini 2.0 Pro (36.0).

# How to access Llama 4

Both Llama 4 Scout and Llama 4 Maverick are now available under Meta's open license. You can download them directly from the official Llama website or via Hugging Face .

To access the models through Meta's services, you can interact with Meta AI on several platforms: WhatsApp, Messenger, Instagram, and Facebook. Currently, access requires logging in with a Meta account, and there is no standalone API endpoint for Meta AI – at least not yet.

If you plan to integrate the models into your own application or infrastructure, please note the licensing terms: If your product or service has more than 700 million monthly active users, you will need to obtain separate permission from Meta. Additionally, the models can be used for research, testing, and most commercial use cases.