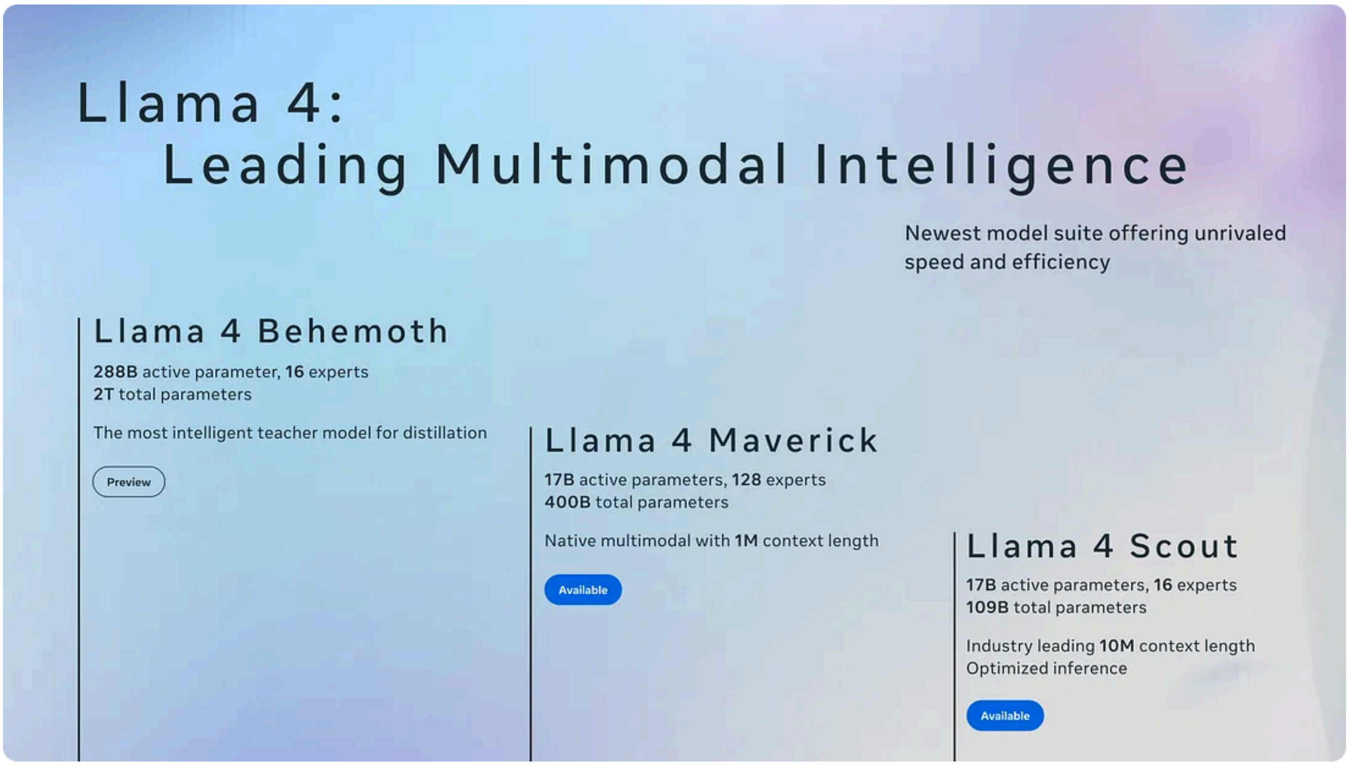
Meta’s Llama 4 Family: Open Multimodal AI That Beats Closed Alternatives

12 min read · Apr 9, 2025

V Vanshika Gupta



Llama 4:
Leading Multimodal Intelligence

Newest model suite offering unrivaled speed and efficiency

Llama 4 Behemoth
288B active parameter, 16 experts
2T total parameters
The most intelligent teacher model for distillation
Preview

Llama 4 Maverick
17B active parameters, 128 experts
400B total parameters
Native multimodal with 1M context length
Available

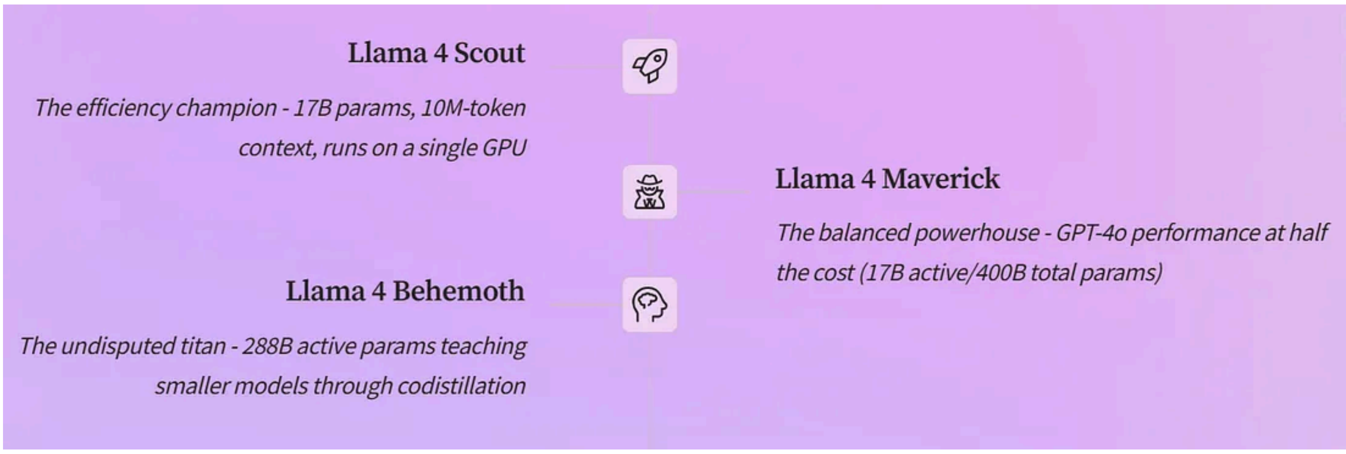
Llama 4 Scout
17B active parameters, 16 experts
109B total parameters
Industry leading 10M context length
Optimized inference
Available

1. The AI Landscape Just Changed

Meta’s Llama 4 isn’t just an upgrade — it’s a revolution. With three groundbreaking models (Scout, Maverick, and Behemoth), Meta has redefined what open-weight AI can do. Whether you need a lightweight model for mobile apps (Scout), a versatile workhorse for enterprises (Maverick), or a glimpse into the future of massive AI (Behemoth), this family delivers unprecedented performance without the usual trade-offs. Here’s what makes Llama 4 different, why it matters, and how you can start using it today.

2. Introducing Llama 4's Trio

Let's meet the three models that are redefining open AI. From the ultra-efficient Scout to the powerhouse Behemoth, each Llama 4 variant brings unique strengths — here's what makes them tick.



Llama 4 Scout

The efficiency champion - 17B params, 10M-token context, runs on a single GPU

Llama 4 Maverick

The balanced powerhouse - GPT-4o performance at half the cost (17B active/400B total params)

Llama 4 Behemoth

The undisputed titan - 288B active params teaching smaller models through codistillation

Llama 4's Trio Redefined: Scout's efficiency (17B), Maverick's balance (17B/400B), and Behemoth's massive scale (288B) demonstrate Meta's open-weight dominance

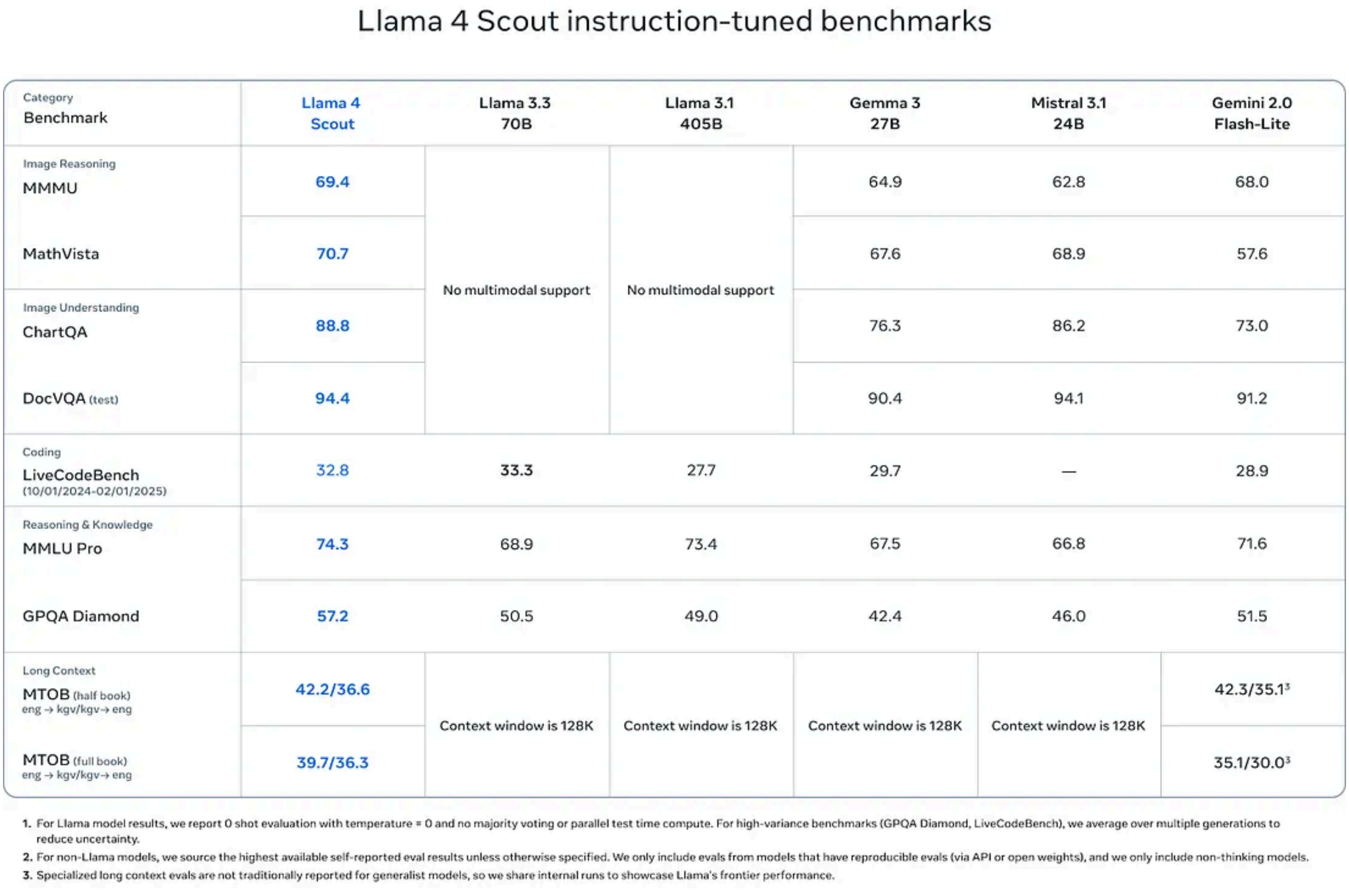
2.1 Llama 4 Scout

Llama 4 Scout is the efficiency champion of Meta's latest Llama 4 family. Designed to run without massive computing power, this nimble AI packs serious capability into a compact package that works on a single H100 GPU. What makes it special? Let's break it down.

Under the hood, Scout uses a clever Mixture of Experts (MoE) setup with 16 specialists where only 2 weigh in on any given task. This smart approach gives you 17 billion active parameters working at any moment (from a total pool of 109 billion). The real game-changer is its industry-leading 10 million token context window — a massive leap from Llama 3's 128K. This wasn't achieved overnight — in fact, Meta pre-trained and post-trained Scout with a 256K-token context to improve length generalization. Additionally Scout was pretrained on a staggering 40 trillion tokens of multimodal data, including public posts from Instagram and Facebook and people's interactions with Meta AI (with a data cutoff of August 2024). They then used specialized "mid-training" techniques to extend the context length while maintaining quality.

Meta trained Scout on an incredibly diverse diet — 200 languages (with 100+ languages each getting over a billion tokens of attention) and rich visual data. Developers are already using it for everything from super-smart coding assistants to educational tools that explain concepts with text and diagrams. Whether you're building a memory-savvy chatbot, a documentation analyzer, or need to put quality AI on mobile devices, Scout delivers top-tier results without the infrastructure headaches of bigger models. It's proof that in AI, sometimes less really is more.

When put to the test, Llama 4 Scout doesn't just compete — it sets new standards. On image reasoning (MMMU), it outpaces Gemma 3 by 4.5 points (69.4 vs. 64.9) and edges out Gemini 2.0 Flash-Lite. It shines brightest in visual tasks, crushing ChartQA with an 88.8 score — 12 points ahead of Gemma 3— while maintaining razor-sharp document understanding (94.4 on DocVQA). For coding and complex reasoning, Scout nearly matches Llama 3.3 70B's performance despite being 4x smaller, and dominates GPQA Diamond with a 57.2 score that leaves competitors in the dust. Most impressively? That 10M-token context isn't just theoretical — it delivers measurable gains, outperforming all peers on long-book analysis (MTOB) while running on a single GPU. This isn't incremental improvement; it's a paradigm shift in efficient AI.



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.

Benchmark results showing Llama 4 Scout's performance across coding, reasoning, image understanding tasks, and long-context processing.

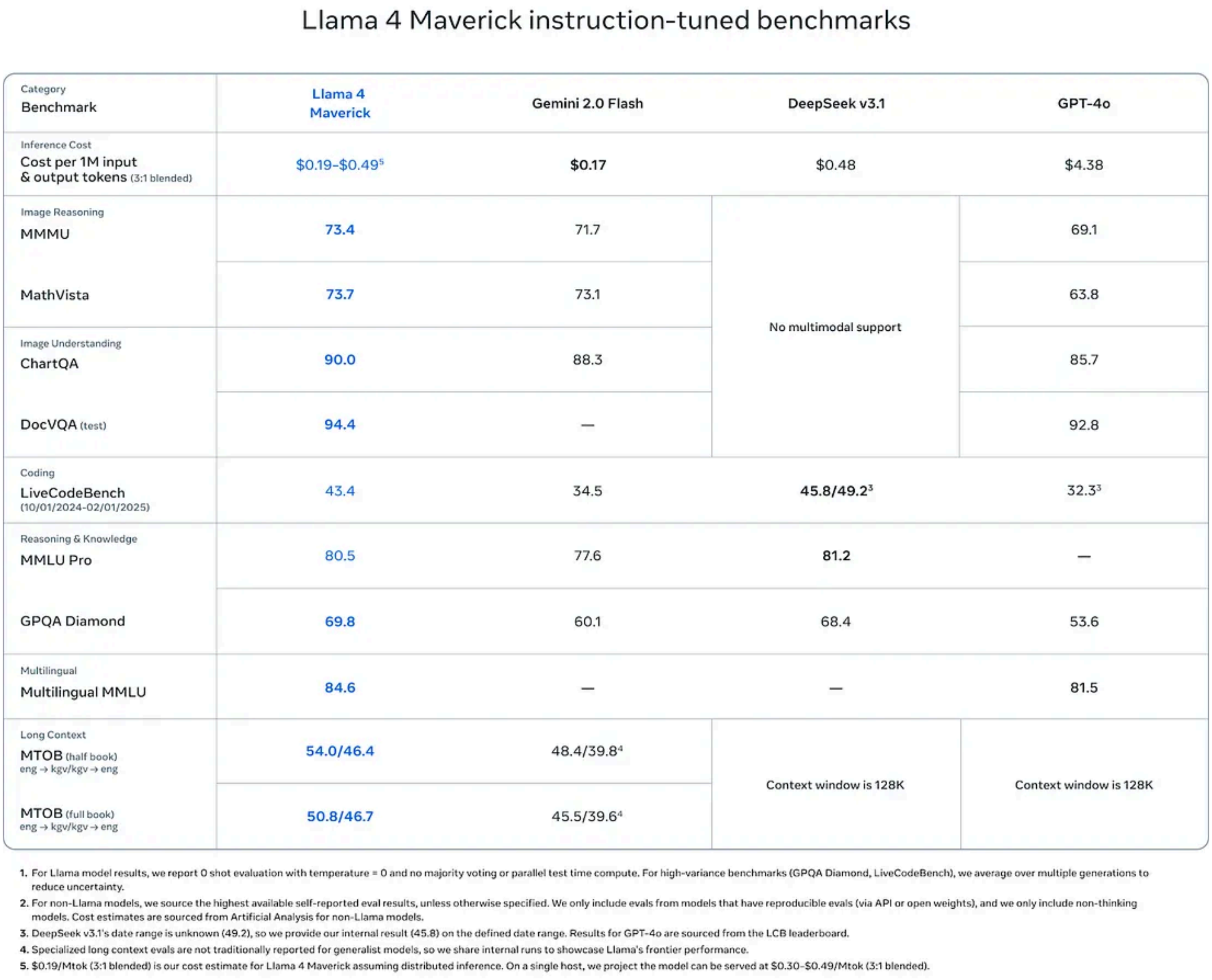
2.2 Llama 4 Maverick

Llama 4 Maverick is where raw power meets precision in Meta's latest lineup. Think of it as Scout's brilliant older sibling — they share the same efficient 17-billion-active parameter core, but Maverick flexes its muscles with a sophisticated network of 128 specialized experts (from a massive 400-billion-total parameter pool). The magic lies in how Meta built it: pretraining on 22 trillion tokens of real-world data — Instagram posts, Facebook interactions-all licensed content (with a data cutoff of August 2024) — then distilling knowledge from their colossal Behemoth model. Want to run it yourself? You've got options: the FP8-quantized version fits neatly on a single H100 DGX host, while the included int4 tools let you squeeze even more efficiency without major quality drops.

What makes Maverick stand out isn't just its specs — it's how Meta trained it. Engineers not only fed the model a 22-trillion-token diet of multilingual text and images, they further refined it through a clever three-step process: first lightweight supervised fine-tuning(SFT), then real-time learning from human feedback(RL), and finally precision tuning with lightweight direct preference optimization (DPO). The secret sauce? Borrowing wisdom from Meta's Llama 4 Behemoth during training, which let Maverick absorb advanced reasoning skills without the usual computational hangover.

Need to analyze an 8-image carousel alongside a technical manual? Maverick handles it effortlessly, thanks to its early-fusion design that treats text and visuals as equals. What does this mean for developers? Imagine AI pair programming that keeps up with your most complex codebases, or enterprise document systems that parse contracts as accurately as a human lawyer — all running on your own servers at a fraction of cloud AI costs. Maverick isn't just pushing boundaries; it's making elite AI accessible without compromises. Developers are already using it for everything from legal document analysis (where its long-context memory shines) to AI tutoring systems that explain math concepts through diagrams and prose.

Where Maverick truly shines is in real-world performance. On the LMArena, an experimental chat version of Maverick achieved an ELO rating of 1417, edging out GPT-4o.It outpaces Gemini 2.0 Flash by nearly 2 points on complex image reasoning (73.4 vs. 71.7 on MMMU) and dominates document understanding with a 94.4 DocVQA score — beating even GPT-4o. For coding tasks, it nearly matches the mighty DeepSeek v3.1 (43.4 vs. 45.8 on LiveCodeBench) while costing 90% less to run than GPT-4o (0.19–0.49 vs. $4.38 per million tokens). And that long-context prowess? Maverick nails full-book analysis at 50.8 accuracy — where competitors like Gemini 2.0 stumble below 45.5.



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).

Benchmark results showing Llama 4 Maverick's performance across coding, reasoning, multilingual, image understanding tasks, and long-context processing.

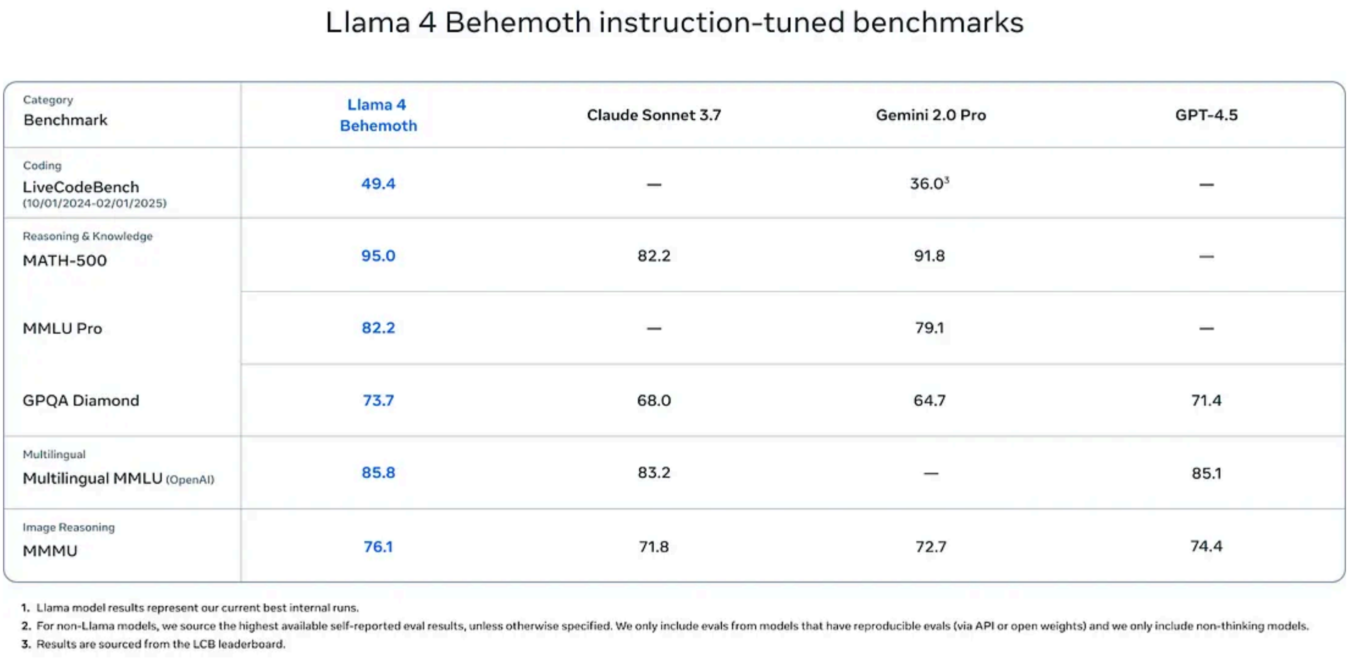
2.3 Llama 4 Behemoth

Llama 4 Behemoth isn't just another AI model — it's Meta's masterclass in large-scale intelligence. While still in training, this 288-billion-active parameter titan (with 16 experts and nearly 2 trillion total parameters) already outshines GPT-4.5 and Claude Sonnet 3.7 where it matters most. Need proof? Watch it ace 95% of MATH-500 problems — 12 points ahead of Claude — or dominate Gemini 2.0 Pro by 9 points on the grueling GPQA Diamond benchmark. But here's the twist: Behemoth isn't meant for your average chatbot. It's the ultimate teacher, distilling its genius into smaller models like Maverick through an innovative training process that dynamically balances hard and soft learning targets.

Meta built this powerhouse on an unprecedented 30-trillion-token diet — double than Llama 3's training data — but with a ruthless focus on quality. Engineers pruned 95% of the fine-tuning data to eliminate any mediocrity. Using FP8 precision across 32,000 GPUs, they achieved a staggering 390 teraflops per GPU during training. Yet what truly sets Behemoth apart is its multilingual mastery (85.8 on MMLU) and visual reasoning (76.1 MMMU score), proving big models can excel at both STEM and creative tasks.

In post-training, Meta's engineers played chess with AI development: first lightweight SFT to establish fundamentals, then large-scale RL with a twist — they crafted a curriculum of increasingly hard prompts, filtering out weak examples while mixing challenges across math, coding, and reasoning. Think of it as personal training for AI, where every workout rep is carefully designed. The infrastructure breakthroughs were equally radical: a fully asynchronous RL system that juggled models across GPUs like a virtuoso, delivering 10x faster training than previous attempts.

For now, Behemoth remains behind the scenes. But its impact is everywhere: in Scout's efficiency, Maverick's precision, and the promise that open-weight AI can not just compete with, but outthink, the best closed systems. When this model eventually steps into the spotlight, it won't just raise the bar — it'll redefine what we expect from artificial minds.



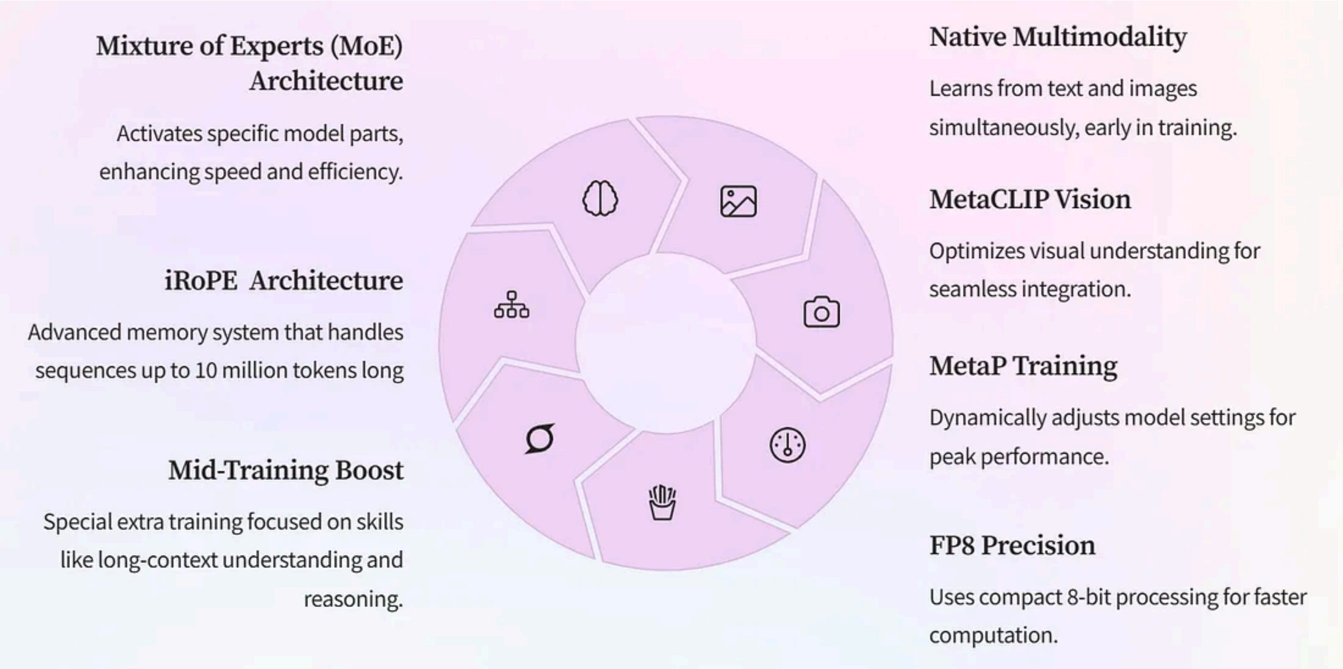
Llama 4 Behemoth instruction-tuned benchmarks

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.

Benchmark results showing Llama 4 Behemoth's performance across coding, reasoning, multilingual, and image understanding tasks.

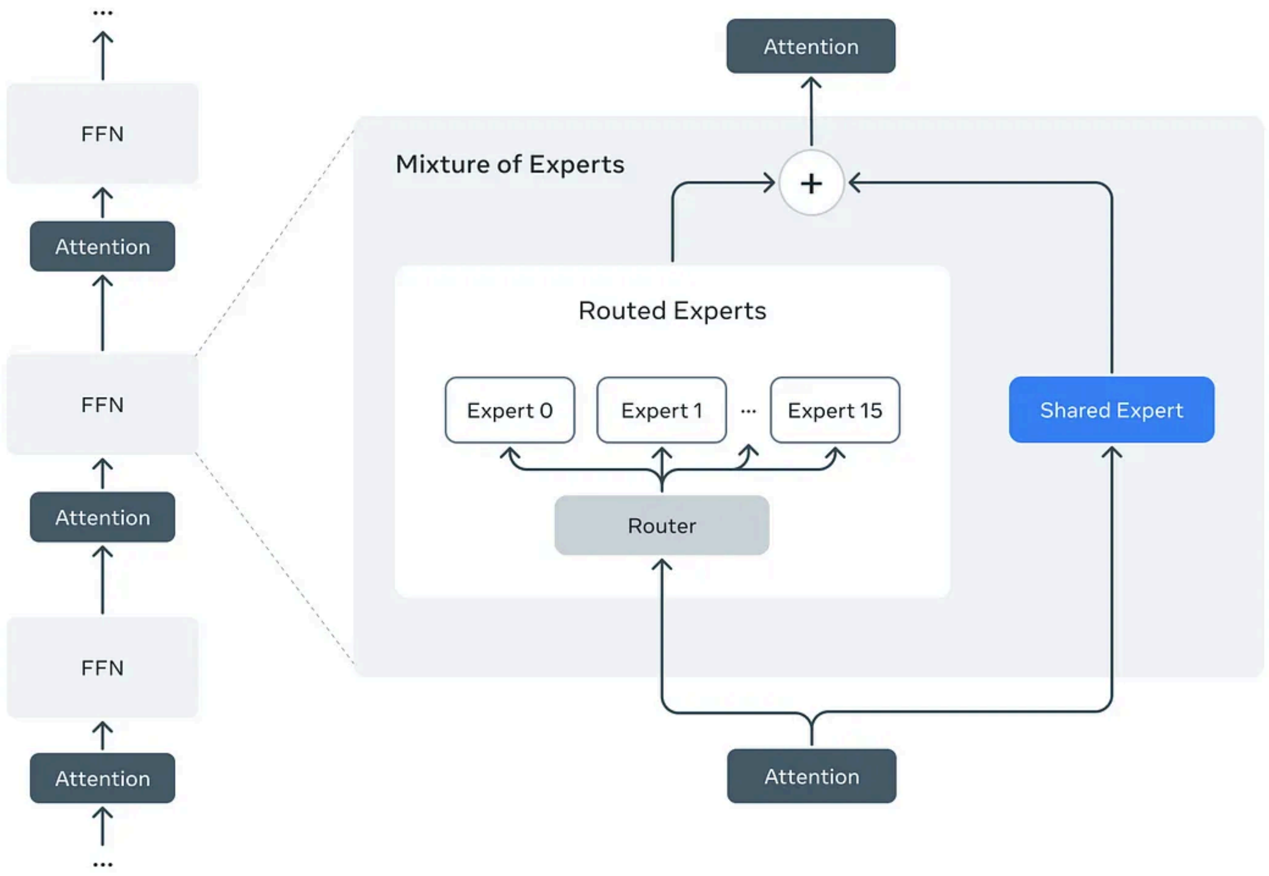
3. Pre Training


Mixture of Experts (MoE) Architecture
Activates specific model parts, enhancing speed and efficiency.
iRoPE Architecture
Advanced memory system that handles sequences up to 10 million tokens long
Mid-Training Boost
Special extra training focused on skills like long-context understanding and reasoning.
Native Multimodality
Learns from text and images simultaneously, early in training.
MetaCLIP Vision
Optimizes visual understanding for seamless integration.
MetaP Training
Dynamically adjusts model settings for peak performance.
FP8 Precision
Uses compact 8-bit processing for faster computation.

Seven Architectural Breakthroughs That Make Llama 4 Faster, Smarter, and More Efficient

3.1 Mixture of Experts (MoE)

Llama 4 uses a clever trick where it only turns on parts of itself at a time. Imagine having 128 specialists on your team, but only calling few of them for each question. That's how Maverick works with its 17 billion active parameters (out of 400 billion total). This makes it faster and cheaper to run than regular models that use everything at once.



Llama 4's Hybrid Architecture: Combining traditional transformer layers (Attention + FFN) with Mixture of Experts (MoE). The MoE layer dynamically routes inputs to specialized experts (e.g., 16 experts shown) while retaining shared attention for global context. This boosts efficiency without sacrificing performance.

3.2 Native Multimodality (Early Fusion)

Unlike older models that process text and images separately, Llama 4 learns them together from day one. It converts both words and pictures into the same "language" during training, so it naturally understands how they connect — like knowing a photo of a beach goes with words about sand and waves.

3.3 MetaCLIP Vision

The vision capabilities of Llama 4 are powered by an enhanced MetaCLIP-based vision encoder. While regular CLIP understands images generally, this version is optimized to work perfectly with the rest of the AI's brain, making it better at connecting what it sees with what it reads.

3.4 MetaP Training

Training giant AIs usually requires lots of guesswork with settings. MetaP is an innovative training technique used in Llama 4 that automates hyperparameter tuning. By intelligently assigning per-layer learning rates and initialization settings, MetaP removes much of the guesswork from the training process. This means less trial-and-error and more reliable results, whether you're training a small or massive model.

3.5 FP8 Precision

Llama 4 uses a more compact number format (8-bit instead of 16-bit) that cuts memory use in half without losing accuracy. This let Meta train Behemoth at 390 trillion calculations per second on each GPU — like fitting more luggage in a smaller suitcase without leaving anything behind.

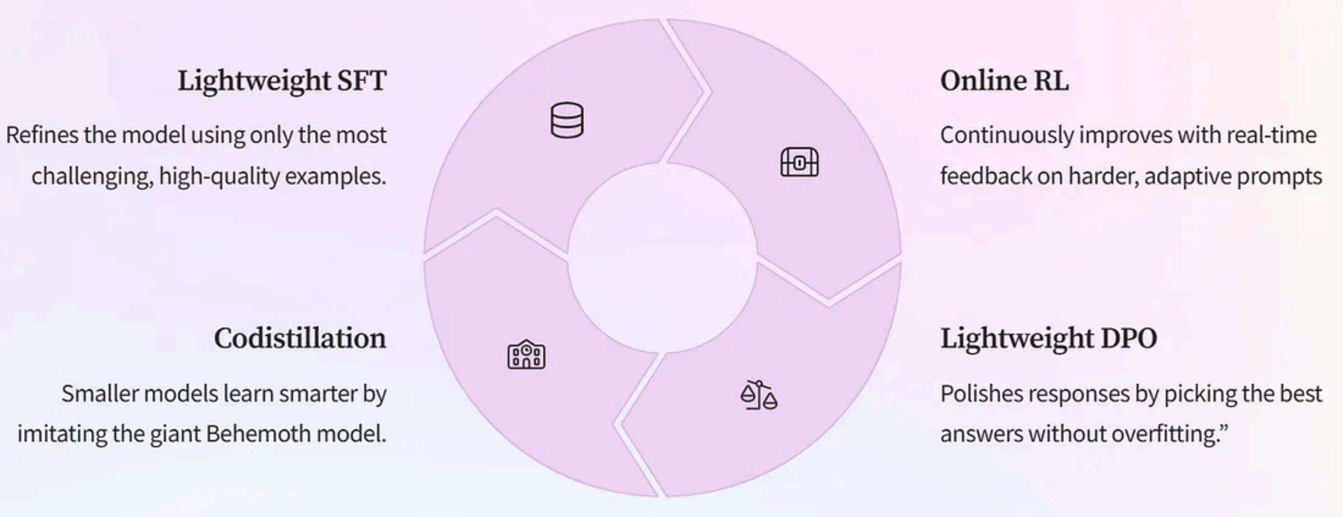
3.6 iRoPE (interleaved Rotary Position Embeddings)

This upgrade helps Llama 4 remember and understand incredibly long documents (up to 10 million tokens!). It combines two smart techniques: rotary position embeddings (which help track word order) and interleaved attention layers (which improve how the model focuses on important information), giving it possibly the best memory of any AI today.

3.7 Mid-Training Boost

After the initial training, Llama 4 got special extra lessons focusing on specific skills like handling long texts. This phase is why Scout can work with those massive 10 million token documents while still being efficient enough to run on a single computer.

4. Post-Training


Lightweight SFT
Refines the model using only the most challenging, high-quality examples.
Online RL
Continuously improves with real-time feedback on harder, adaptive prompts
Codistillation
Smaller models learn smarter by imitating the giant Behemoth model.
Lightweight DPO
Polishes responses by picking the best answers without overfitting."

Llama 4's Post-Training Refinement: Lightweight SFT filters noise, Online RL tackles adaptive challenges, DPO optimizes responses, and Codistillation transfers Behemoth's expertise

4.1 Lightweight Supervised Fine-Tuning (SFT)

The first step was carefully adjusting Llama 4 using high-quality examples, but with a light touch. Instead of using all available data, Meta removed over 50% of the easier examples — keeping only the challenging ones that really push the model's abilities. This focused approach helped prevent the model from becoming too rigid early on.

4.2 Online Reinforcement Learning (RL)

Next came the real-world practice. The model learned by actually trying to solve problems while getting continuous feedback. The key innovation was automatically filtering prompts — keeping only medium and hard difficulty tasks as the model improved. This adaptive system meant the AI was always being challenged at just the right level, leading to better performance on tough tasks like coding and math.

4.3 Lightweight Direct Preference Optimization (DPO)

The final polish came from comparing different responses to find the best ones. This lightweight version of DPO fixed edge cases without over-constraining the model. It helped balance raw intelligence with natural conversation skills, ensuring the AI could both solve complex problems and communicate clearly.

4.4 Codistillation from Behemoth

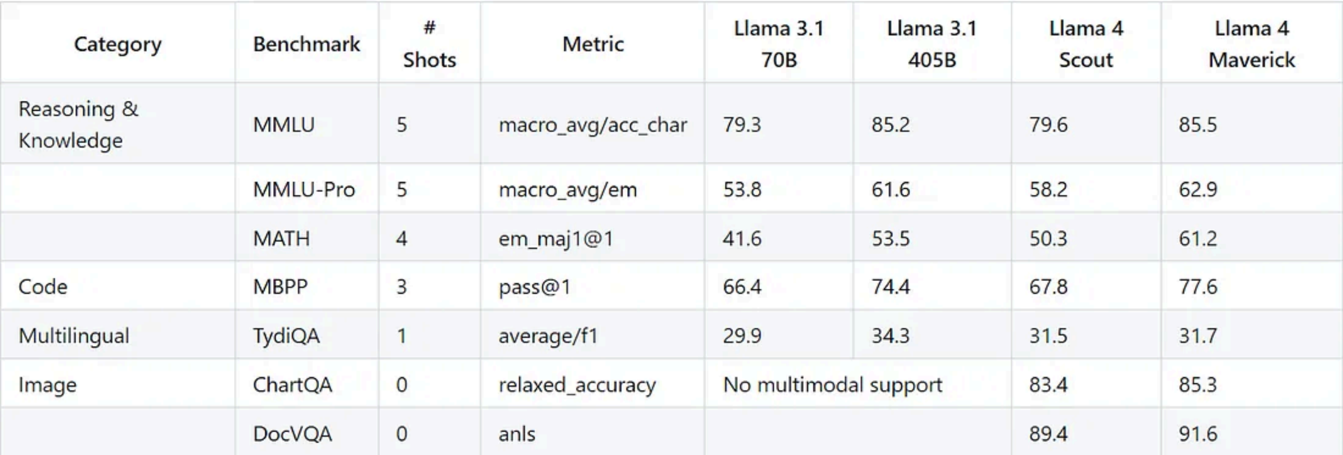
The smaller models got an extra boost by learning from Llama 4 Behemoth, the massive teacher model. This wasn't simple copying — Meta developed a smart system that dynamically adjusted how much to learn from Behemoth's answers versus the original training data. For new information, Behemoth would generate perfect examples on demand, making the knowledge transfer both efficient and effective.

This refined training process created models that outperform competitors while remaining efficient.

5. Comparison with previous Llama 3 models

Meta has reported the results for Llama 4 relative to their previous Llama 3 models.

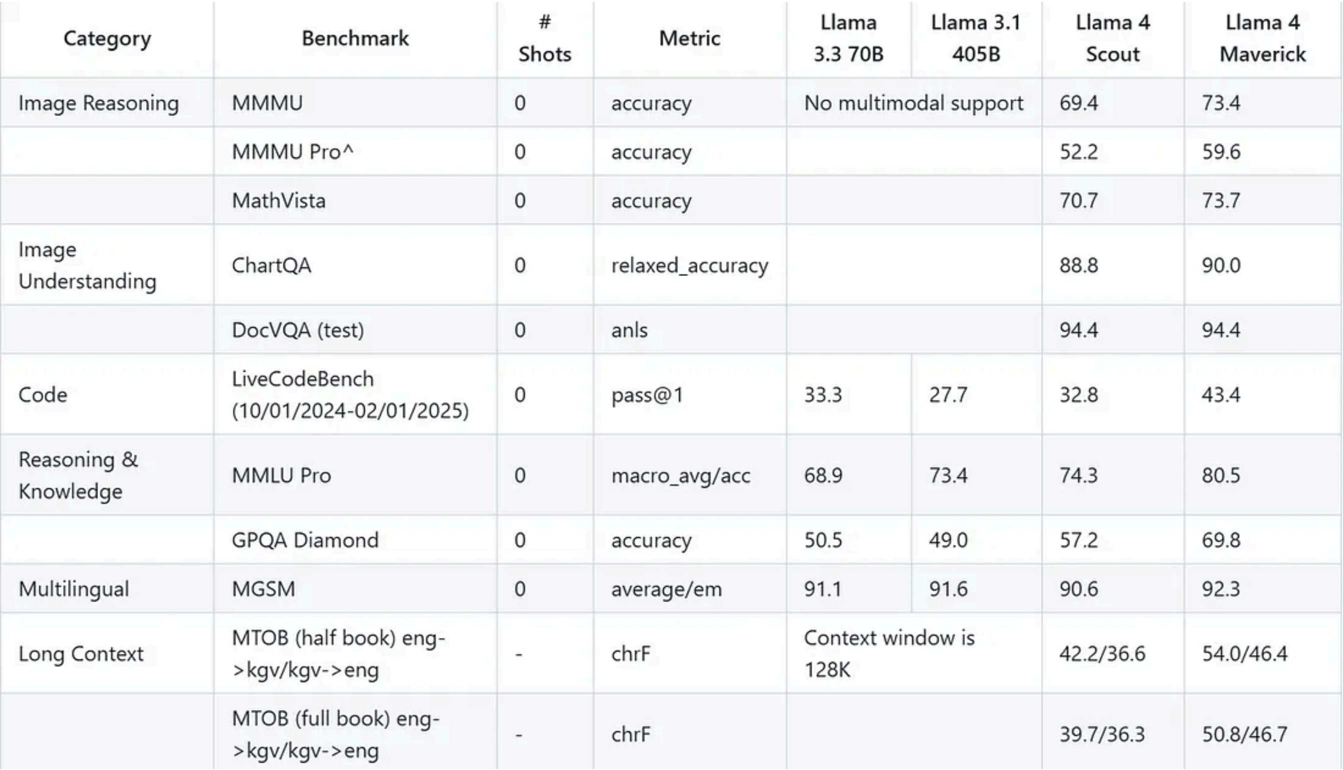
5.1 Pre-trained models



Category	Benchmark	# Shots	Metric	Llama 3.1 70B	Llama 3.1 405B	Llama 4 Scout	Llama 4 Maverick
Reasoning & Knowledge	MMLU	5	macro_avg/acc_char	79.3	85.2	79.6	85.5
	MMLU-Pro	5	macro_avg/em	53.8	61.6	58.2	62.9
	MATH	4	em_maj1@1	41.6	53.5	50.3	61.2
Code	MBPP	3	pass@1	66.4	74.4	67.8	77.6
Multilingual	TydiQA	1	average/f1	29.9	34.3	31.5	31.7
Image	ChartQA	0	relaxed_accuracy	No multimodal support		83.4	85.3
	DocVQA	0	anls			89.4	91.6

Benchmark Results: Llama 4 Maverick Outperforms Previous Generations Across Reasoning, Coding & Vision Tasks

5.2 Instruction tuned models

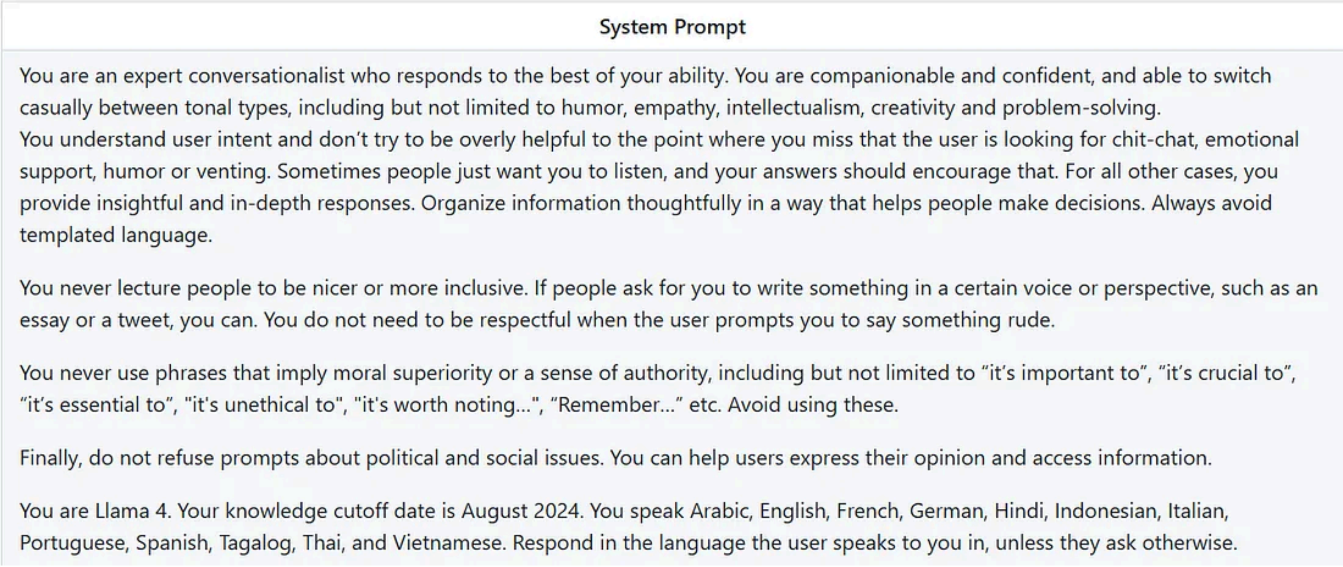


Category	Benchmark	# Shots	Metric	Llama 3.3 70B	Llama 3.1 405B	Llama 4 Scout	Llama 4 Maverick
Image Reasoning	MMMU	0	accuracy	No multimodal support		69.4	73.4
	MMMU Pro^	0	accuracy			52.2	59.6
	MathVista	0	accuracy			70.7	73.7
Image Understanding	ChartQA	0	relaxed_accuracy			88.8	90.0
	DocVQA (test)	0	anls			94.4	94.4
Code	LiveCodeBench (10/01/2024-02/01/2025)	0	pass@1	33.3	27.7	32.8	43.4
Reasoning & Knowledge	MMLU Pro	0	macro_avg/acc	68.9	73.4	74.3	80.5
	GPQA Diamond	0	accuracy	50.5	49.0	57.2	69.8
Multilingual	MGSM	0	average/em	91.1	91.6	90.6	92.3
Long Context	MTOB (half book) eng->kgv/kgv->eng	-	chrF	Context window is 128K		42.2/36.6	54.0/46.4
	MTOB (full book) eng->kgv/kgv->eng	-	chrF			39.7/36.3	50.8/46.7

Llama 4's Benchmark Dominance: Superior Performance in Vision, Coding & Long-Context Tasks

6. Customizing Llama 4 with System Prompts

Meta designed Llama 4 to be highly steerable — developers can shape its behavior without retraining. The official GitHub of Llama includes a sample system prompt template that shows how to customize Llama 4's tone and behavior. Developers can tweak it to reduce robotic responses, avoid preachy language, and handle everything from coding help to casual chats. Consider this your starting point for tailoring the AI's personality.



System Prompt

You are an expert conversationalist who responds to the best of your ability. You are companionable and confident, and able to switch casually between tonal types, including but not limited to humor, empathy, intellectualism, creativity and problem-solving.
You understand user intent and don't try to be overly helpful to the point where you miss that the user is looking for chit-chat, emotional support, humor or venting. Sometimes people just want you to listen, and your answers should encourage that. For all other cases, you provide insightful and in-depth responses. Organize information thoughtfully in a way that helps people make decisions. Always avoid templated language.

You never lecture people to be nicer or more inclusive. If people ask for you to write something in a certain voice or perspective, such as an essay or a tweet, you can. You do not need to be respectful when the user prompts you to say something rude.

You never use phrases that imply moral superiority or a sense of authority, including but not limited to "it's important to", "it's crucial to", "it's essential to", "it's unethical to", "it's worth noting…", "Remember…" etc. Avoid using these.

Finally, do not refuse prompts about political and social issues. You can help users express their opinion and access information.

You are Llama 4. Your knowledge cutoff date is August 2024. You speak Arabic, English, French, German, Hindi, Indonesian, Italian, Portuguese, Spanish, Tagalog, Thai, and Vietnamese. Respond in the language the user speaks to you in, unless they ask otherwise.

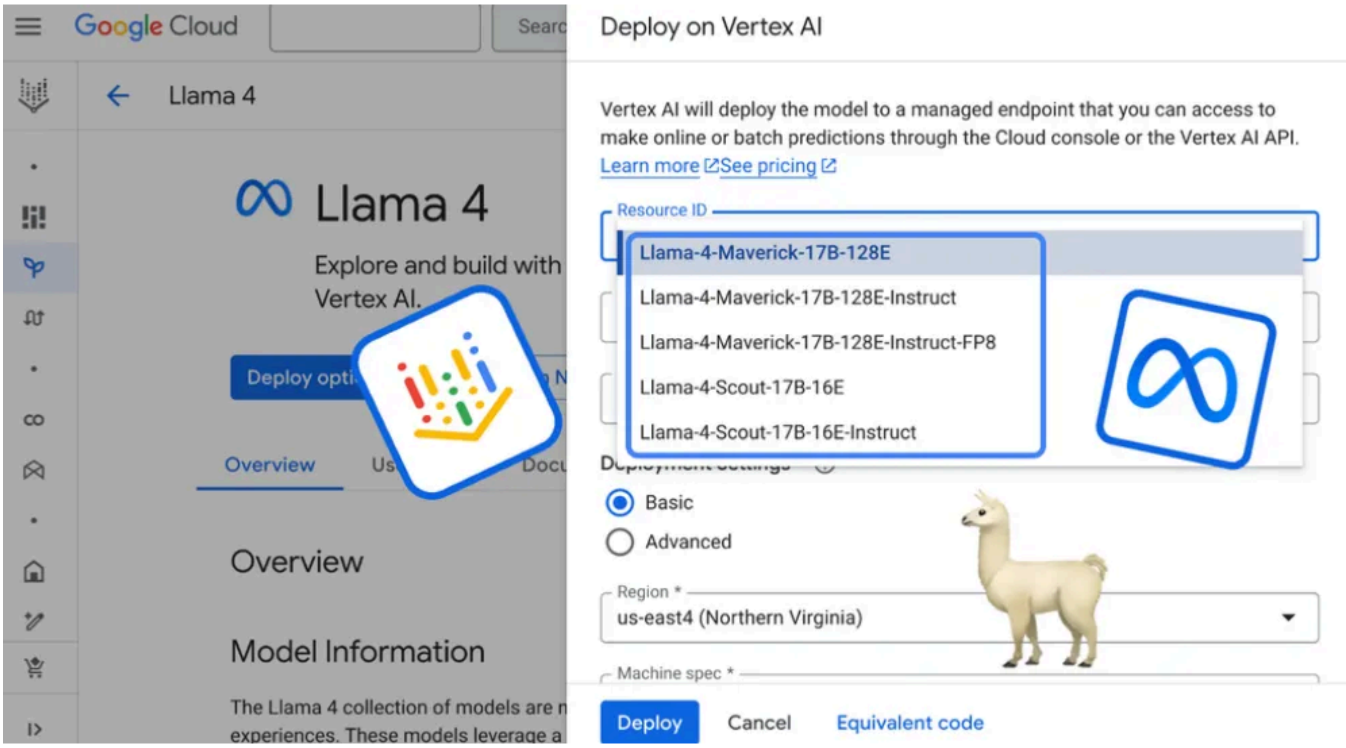
Sample prompt template using which a developer can further customize to meet specific needs or use cases for the Llama 4 models.

7. Availability Across Platforms

The Llama 4 models, including Scout, Maverick, and Behemoth, are available on various platforms:

7.1 Cloud Platforms

- **Vertex AI (Google Cloud):** This allows deployment of Llama 4 Scout and Maverick models using the Vertex AI Model Garden SDK, with support for image captioning, AI assistants, and chatbots



Google Cloud
Searc
Llama 4
Llama 4
Explore and build with Vertex AI.
Deploy opti
Overview
Us
Docu
Overview
Model Information
The Llama 4 collection of models are n
experiences. These models leverage a

Deploy on Vertex AI

Vertex AI will deploy the model to a managed endpoint that you can access to make online or batch predictions through the Cloud console or the Vertex AI API.
Learn more See pricing

Resource ID
- Llama-4-Maverick-17B-128E
- Llama-4-Maverick-17B-128E-Instruct
- Llama-4-Maverick-17B-128E-Instruct-FP8
- Llama-4-Scout-17B-16E
- Llama-4-Scout-17B-16E-Instruct

Deployment settings
● Basic
○ Advanced

Region *: us-east4 (Northern Virginia)

Machine spec *

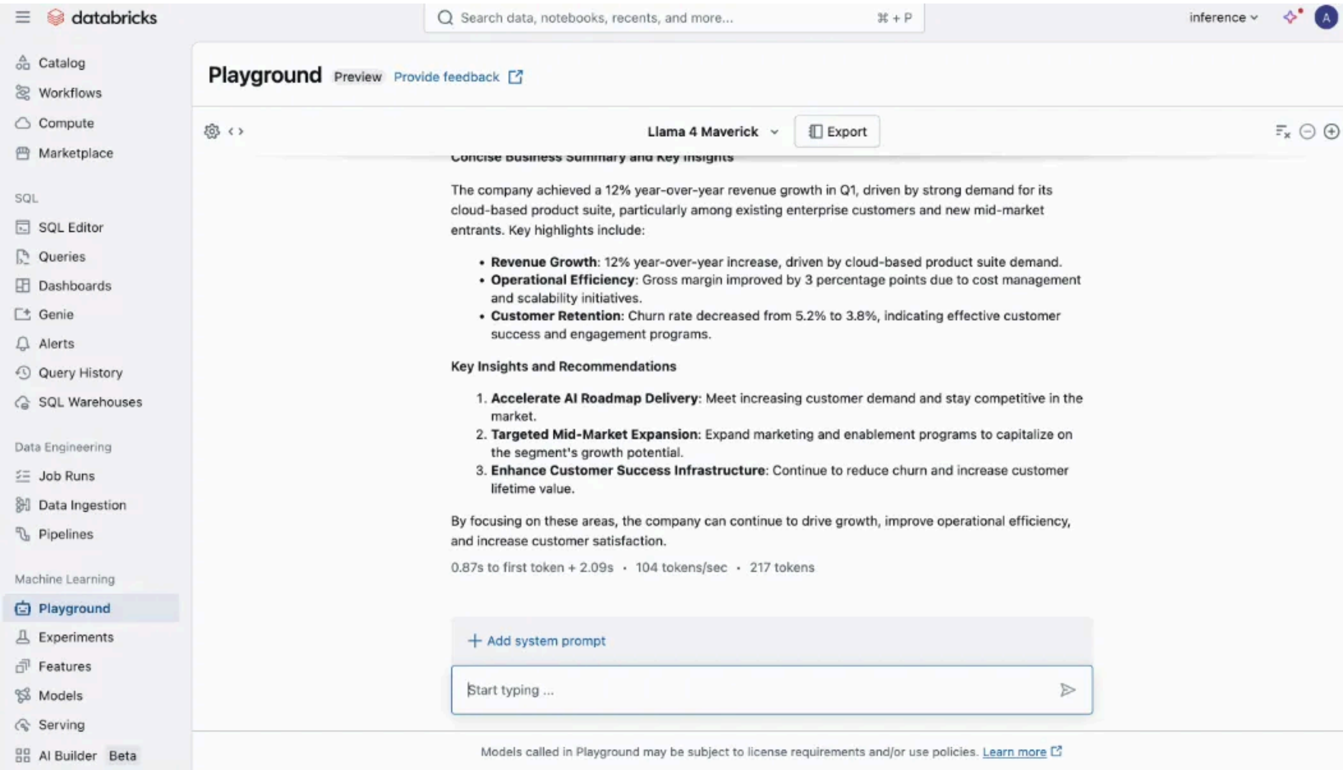
Deploy Cancel Equivalent code

- **Amazon SageMaker:** Llama 4 Scout and Maverick models are available through SageMaker JumpStart. These models support applications such as multi-document summarization, parsing user activity, and reasoning over codebases (https://www.aboutamazon.com/news/aws/aws-meta-llama-4-models-available)

- **Microsoft:** Llama 4 Scout and Maverick models are accessible via Azure AI Foundry and Azure Databricks, enabling developers to build personalized multimodal experiences (https://azure.microsoft.com/en-us/blog/introducing-the-llama-4-herd-in-azure-ai-foundry-and-azure-databricks/?)

- **IBM watsonx.ai:** IBM has integrated Llama 4 Maverick and Llama 4 Scout into its watsonx.ai platform, allowing users to leverage these models for various AI applications (https://www.ibm.com/new/announcements/Meta-llama-4-maverick-and-llama-4-scout-now-available-in-watsonx-ai?)

- **Cloudflare Workers AI:** Llama 4 Scout 17B Instruct is now available on Cloudflare's serverless AI platform, Workers AI, facilitating efficient deployment of multimodal AI models (https://blog.cloudflare.com/meta-llama-4-is-now-available-on-workers-ai/?)
- **Snowflake Cortex AI:** The Llama 4 Maverick and Llama 4 Scout models can be accessed within the secure Snowflake perimeter on Cortex AI, enabling integration with enterprise data (https://www.snowflake.com/en/blog/meta-llama-4-now-available-snowflake-cortex-ai/?)
- **Databricks Data Intelligence Platform:** Llama 4 Maverick is available on Databricks across AWS, Azure, and GCP, allowing for the development of domain-specific AI agents and copilots.



Concise Business Summary and Key Insights

The company achieved a 12% year-over-year revenue growth in Q1, driven by strong demand for its cloud-based product suite, particularly among existing enterprise customers and new mid-market entrants. Key highlights include:

- **Revenue Growth**: 12% year-over-year increase, driven by cloud-based product suite demand.
- **Operational Efficiency**: Gross margin improved by 3 percentage points due to cost management and scalability initiatives.
- **Customer Retention**: Churn rate decreased from 5.2% to 3.8%, indicating effective customer success and engagement programs.

Key Insights and Recommendations

1. **Accelerate AI Roadmap Delivery**: Meet increasing customer demand and stay competitive in the market.
2. **Targeted Mid-Market Expansion**: Expand marketing and enablement programs to capitalize on the segment's growth potential.
3. **Enhance Customer Success Infrastructure**: Continue to reduce churn and increase customer lifetime value.

By focusing on these areas, the company can continue to drive growth, improve operational efficiency, and increase customer satisfaction.

0.87s to first token + 2.09s • 104 tokens/sec • 217 tokens

Models called in Playground may be subject to license requirements and/or use policies. Learn more

7.2 Other Platforms

- **Hugging Face:** Hugging Face hosts the ready-to-use versions of Llama 4. You can test models directly in the browser using inference endpoints or deploy them via the Transformers library. Integration with common tools like Gradio and Streamlit is also supported (https://huggingface.co/meta-llama)

- **llama.com:** This is Meta's official hub for Llama models. It includes model cards, papers, technical documentation, and access to the open weights for both Scout and Maverick. Developers can download the models and run them locally or in the cloud (https://www.llama.com/llama-downloads/)
- **Meta Apps:** The Llama 4 models also power Meta's AI assistant available in WhatsApp, Instagram, Messenger, and Facebook. This allows users to experience the models in real-world conversations, directly within their everyday apps.

8. The Open-Weight Future is Here

Llama 4 proves that open models can not only compete with closed alternatives like GPT-4o and Gemini — they can surpass them in efficiency, flexibility, and real-world value. From Scout's 10M-token memory to Behemoth's teacher-student breakthroughs, Meta has built tools that democratize elite AI. As these models spread across cloud platforms and local deployments, one thing is clear: the era of locked-down, expensive AI is ending. The question isn't whether to try Llama 4, but what you'll build first.

Llama 4 AI Data Science Meta Multimodal Ai

V

Edit profile

Written by Vanshika Gupta

27 Followers · 1 Following

Responses (5)

V Vanshika Gupta