Llama 4 Benchmark and Model Comparison Report

Llama 4: Leading intelligence. Unrivaled speed and efficiency.

The most accessible and scalable generation of Llama is here. Native multimodality, mixture-of-experts models, super long context windows, step changes in performance, and unparalleled efficiency—all in easy-to-deploy sizes custom fit for how you want to use it.

Model Cards

Llama 4 Scout A class-leading natively multimodal model that offers superior text and visual intelligence, efficient single H100 GPU performance, and a 10M context window for seamless long document analysis.

Llama 4 Maverick An industry-leading multimodal model for image and text understanding that delivers groundbreaking intelligence and fast responses at a low cost.

Llama 4 Behemoth Preview An early preview (it's still training!) of the Llama 4 teacher model used to distill Llama 4 Scout and Llama 4 Maverick.

Key Features

- **Natively Multimodal** — Llama 4 models leverage early fusion by pre-training on large amounts of unlabeled text and vision tokens, marking a significant step forward from separate, frozen multimodal weights.
- **Advanced Problem Solving** — Both Llama 4 Scout and Llama 4 Maverick tackle intricate problems, offering intelligent solutions across complex domains.
- **Unparalleled Long Context** — With Llama 4 Scout supporting up to 10M tokens of context (the longest available in the industry), new use cases in memory, personalization, and multimodal applications become possible.
- **Expert Image Grounding** — These models excel in aligning user prompts with relevant visual concepts, anchoring responses to specific image regions.
- **Multilingual Writing** — Pre-trained and fine-tuned for robust text understanding across 12 languages, Llama 4 supports global development and deployment.

Individual Model Benchmark Tables

Gemini 2.5 Pro Experimental 03-25 (Source: https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#enhanced-reasoning)

Table 1: Benchmark Comparison Featuring Gemini 2.5 Pro Exp. 03-25

Benchmark	Gemini 2.5 Pro	o3-mini	GPT-4.5	Claude 3.7	Grok 3 Beta	DeepSeek R1
Reasoning & knowledge						
Humanity's Last Exam (no tools)	18.8	14.0*	6.4	8.9	–	8.6*
Science						
GPQA diamond (single, pass@1)	84.0	79.7	71.4	78.2	80.2	71.5
GPQA diamond (multiple)	–	–	–	**84.8**	**84.6**	–
Mathematics						
AIME 2025 (single, pass@1)	**86.7**	86.5	–	49.5	77.3	70.0
AIME 2025 (multiple)	–	–	–	–	**93.3**	–
AIME 2024 (single, pass@1)	**92.0**	87.3	36.7	61.3	83.9	79.8
AIME 2024 (multiple)	–	–	–	**80.0**	**93.3**	–
Code generation						
LiveCodeBench v5 (single, pass@1)	70.4	**74.1**	–	–	70.6	64.3
LiveCodeBench v5 (multiple)	–	–	–	–	**79.4**	–
Code editing						
Aider Polyglot	74.0 / 68.6	60.4[d]	44.9[d]	64.9[d]	–	56.9[d]
Agentic coding						
SWE-bench verified	63.8	49.3	38.0	**70.3**	–	49.2
Factuality						
SimpleQA	**52.9**	13.8	62.5	–	43.6	30.1
Visual reasoning						
MMMU (single)	**81.7**	No MM†	74.4	75.0	76.0	No MM†
MMMU (multiple)	–	–	–	–	**78.0**	–
Image understanding						
Vibe-Eval (Reka)	69.4	No MM†	–	–	–	No MM†
Long context						
MRCR (128k avg)	**94.5**	61.4	64.0	–	–	–
MRCR (1M pointwise)	**83.1**	–	–	–	–	–
Multilingual performance						
Global MMLU (Lite)	**89.8**	–	–	–	–	–

* Text problems only.
[d] Diff performance. † No multimodal support reported/applicable. pass@1: Single attempt. Multiple: Multiple attempts/voting. Gemini results: default sampling (pass@1), model `gemini-2.5-pro-exp-03-25`. Non-Gemini results: self-reported. Sources include https://agi.safe.ai/, https://matharena.ai/, https://livecodebench.github.io/, https://aider.chat/docs/leaderboards. Source for Gemini table: https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#enhanced-reasoning.

Llama 4 Maverick (Source: https://www.llama.com)

Table 2: Llama 4 Maverick Benchmark Comparison

Category / Benchmark	Llama 4 Maverick	Gemini 2.0 Flash	DeepSeek v3.1	GPT-4o
Inference Cost ($/1M tokens In/Out)				
	$0.19–0.49[a]	$0.17	$0.48	$4.38
Image Reasoning				
MMMU	73.4	71.7	No MM[†]	69.1
MathVista	73.7	73.1	No MM[†]	63.8
Image Understanding				
ChartQA	**90.0**	88.3	–	85.7
DocVQA (test)	**94.4**	–	–	92.8
Coding				
LiveCodeBench (10/24–02/25)	**43.4**	34.5	**45.8 / 49.2**[b]	32.3
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)	**54.0 / 46.4**	48.4 / 39.8	128k context[‡]	128k context[‡]
MTOB (full book)	**50.8 / 46.7**	45.5 / 39.6	128k context[‡]	128k context[‡]

[a] $0.19/1Mtok (3:1 blended) estimated distributed inference cost. [b] DeepSeek v3.1 internal result (45.8) used as range unknown. [†] No multimodal support reported/applicable. [‡] Context window limits reported result. Llama results: 0-shot, temp=0, averaged for high-variance. Non-Llama: highest self-reported. Cost estimates (non-Llama): Artificial Analysis. Source: https://www.llama.com

Llama 4 Scout (Source: https://www.llama.com)

Table 3: Llama 4 Scout Benchmark Comparison

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	–	–	64.9	62.8	68.6
MathVista	70.7	–	–	67.6	68.9	57.6
Image Understanding						
ChartQA	88.8	No MM†	No MM†	76.3	**86.2**	73.0
DocVQA	**94.4**	–	–	90.4	**94.1**	91.2
Coding						
LiveCodeBench (10/24–02/25)	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)	**42.2 / 36.6**	128k context‡	128k context‡	128k context‡	128k context‡	42.3 / 35.1
MTOB (full book)	**39.7 / 36.3**	–	–	–	–	35.1 / 30.0

† No multimodal support reported/applicable. ‡ Context window limits reported result. Llama results: 0-shot, temp=0, averaged for high-variance. Non-Llama: highest self-reported. Source: https://www.llama.com

Llama 4 Behemoth (Source: https://www.llama.com)

Table 4: Llama 4 Behemoth Benchmark Comparison (Preview)

Category / Benchmark	Llama 4 Behemoth	Claude Sonnet 3.7	Gemini 2.0 Pro	GPT-4.5
Coding				
LiveCodeBench (10/24–02/25)	**49.4**	–	36.0	–
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

Llama results: Current best internal runs (preview model). Non-Llama: highest self-reported. Source: https://www.llama.com

Combined Model Benchmark Comparison

Table 5: Combined Model Benchmark Comparison (GPQA Diamond & MMLU)

Model	GPQA Diamond	MMLU / Global MMLU
Gemini 2.5 Pro Exp03-25	84.0	89.8
o3-mini	79.7	–
GPT-4.5	71.4	–
Claude 3.7	78.2	–
Grok 3 Beta	80.2	–
DeepSeek R1	71.5	–
Llama 4 Maverick	69.8	80.5
Gemini 2.0 Flash	60.1	–
DeepSeek v3.1	68.4	–
GPT-4o	53.6	–
Llama 4 Scout	57.2	74.3
Llama 3.3 70B	50.5	–
Llama 3.1 405B	49.0	–
Gemma 3 (27B)	42.4	–
Mistral 3.1 (24B)	46.0	–
Gemini 2.0 Flash-Lite	51.5	–
Llama 4 Behemoth	73.7	82.2
Claude Sonnet 3.7	68.0	–
Gemini 2.0 Pro	64.7	–

GPQA Diamond: Single attempt (pass@1).
MMLU / Global MMLU: MMLU Pro or Global MMLU (Lite) where available.