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.

Benchmark & Model Comparison Tables

Gemini Table

Benchmark	**Gemini 2.5 Pro** (Experimental 03-25)	**OpenAI o3-mini** (High)	**OpenAI GPT-4.5**	**Claude 3.7 Sonnet** (64k Extended Thinking)	**Grok 3** (Extende
Reasoning & knowledge					
Humanity's Last Exam (no tools)	18.8%	14.0%*	6.4%	8.9%	–
Science					
GPQA diamond (single attempt, pass@1)	84.0%	79.7%	71.4%	78.2%	80.2%
GPQA diamond (multiple attempts)	–	–	–	**84.8%**	**84.6%**
Mathematics					
AIME 2025 (single attempt, pass@1)	**86.7%**	86.5%	–	49.5%	77.3%
AIME 2025 (multiple attempts)	–	–	–	–	**93.3%**
AIME 2024 (single attempt, pass@1)	**92.0%**	87.3%	36.7%	61.3%	83.9%
AIME 2024 (multiple attempts)	–	–	–	**80.0%**	**93.3%**
Code generation					
LiveCodeBench v5 (single attempt, pass@1)	70.4%	**74.1%**	–	–	70.6%
LiveCodeBench v5 (multiple attempts)	–	–	–	–	**79.4%**
Code editing					
Aider Polyglot	74.0% / 68.6%	60.4% (diff)	44.9% (diff)	64.9% (diff)	–
Agentic coding					
SWE-bench verified	63.8%	49.3%	38.0%	**70.3%**	–
Factuality					
SimpleQA	**52.9%**	13.8%	62.5%	–	43.6%
Visual reasoning					
MMMU (single attempt)	**81.7%**	No MM support	74.4%	75.0%	76.0%
MMMU (multiple attempts)	No MM support	–	–	–	**78.0%**
Image understanding					
Vibe-Eval (Reka)	69.4%	No MM support	–	–	–
Long context					
MRCR (128k average)	**94.5%**	61.4%	64.0%	–	–
MRCR (1M pointwise)	**83.1%**	–	–	–	–
Multilingual performance					
Global MMLU (Lite)	**89.8%**	–	–	–	–

Footnotes:

- * indicates evaluation on **text problems only** (no images).
- "diff" = performance difference from base output after edits (for Aider Polyglot).
- "pass@1" = first-attempt success rate (no majority vote).

Methodology & Sources:

- **Gemini results:** Run with default sampling (pass@1) using the model-id `gemini-2.5-pro-exp-03-25` on AI Studio API. Multiple trials are averaged to reduce variance.
- **Non-Gemini results:** Sourced from providers' self-reported numbers and official reports.
- **Result sources:**
  - Humanity's Last Exam: `https://agi.safe.ai/` | `https://scale.com/leaderboard/humanitys_last_exam`
  - AIME 2025: `https://matharena.ai/`
  - LiveCodeBench: `https://livecodebench.github.io/`
  - Aider Polyglot: `https://aider.chat/docs/leaderboards`

Llama Table 1

Category / Benchmark	**Llama 4 Maverick**	**Gemini 2.0 Flash**	**DeepSeek v3.1**	**GPT-4o**
Inference Cost				
Price per 1M Input & Output tokens	**$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	(No multimodal support)	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**[2]	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)	**54.0 / 46.4**	48.4 / 39.8[0]	(Context window is 128K)	(128K)
MTOB (full book)	**50.8 / 46.7**	45.5 / 39.6[1]	(Context window is 128K)	(128K)

Footnotes:

1. Llama model results are 0-shot with temperature = 0; high-variance benchmarks are averaged over multiple generations.
2. For non-Llama models, highest available self-reported eval results are shown from reproducible evaluations.
3. Cost estimates for non-Llama models are from Artificial Analysis.
4. DeepSeek v3.1's internal result (45.8) is used as its range is unknown.
5. **$0.19/1Mtok (3:1 blended)** represents the distributed inference cost estimate for Llama 4 Maverick.

Llama Table 2

Category / Benchmark	**Llama 4 Scout**	**Llama 3.3 70B**	**Llama 3.1 405B**	**Gemma 3 (27B)**	**Mistra (24B)**
Image Reasoning					
MMMU	69.4	–	–	64.9	62.8
MathVista	70.7	–	–	67.6	68.9
Image Understanding					
ChartQA	88.8	No multimodal support	No multimodal support	76.3	**86.2**
DocVQA	**94.4**	–	–	90.4	**94.1**
Coding					
LiveCodeBench (10/01/2024–02/01/2025)	32.8	**33.3**	27.7	29.7	–
Reasoning & Knowledge					
MMLU Pro	74.3	68.9	73.4	67.5	66.8
GPQA Diamond	**57.2**	50.5	49.0	42.4	46.0
Long Context					
MTOB (half book)	**42.2 / 36.6**	(Context window is 128K)	(Context window is 128K)	(Context window is 128K)	(Contex 128K)
MTOB (full book)	**39.7 / 36.3**	–	–	–	–

Footnotes:

1. Llama model results are reported 0-shot with temperature = 0; averaging is applied for high-variance benchmarks.
2. For non-Llama models, results are the highest available self-reported evaluations from reproducible sources.

Llama Table 3

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	–
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

Footnotes:

1. Llama model results represent the current best internal runs.
2. For non-Llama models, evaluation results are sourced from reproducible self-reported data.