Benchmark / Category	L4 Behemoth	Llama4 Maverick	Llama4 Scout	Llama 3.1 405B	Llama 3.3 70B	Gemini 2.5 Pro	Gemini 2.0 Pro	Gemini 2.0 Flash	Gemini 2.0 Flash-Lite	GPT-4.5	GPT-4o	Claude 3.7	Claude Sonnet 3.7	Grok 3 Beta	DeepSeek R1	DeepSeek v3.1	Gemma 3 (27B)	Mistral 3.1 (24B	o3-mini
Inference Cost ($/1M tokens In/Out)																			
	–	$0.19–0.49[a]	–	–	–	–	–	$0.17	–	–	$4.38	–	–	–	$0.48	–	–	–	–
Reasoning & Knowledge																			
Humanity's Last Exam (no tools)	–	–	–	–	–	18.8	–	–	–	6.4	–	8.9	–	–	8.6[b]	–	–	–	14.0[b]
MMLU Pro	**82.2**	**80.5**	74.3	73.4	**81.2**	–	79.1	77.6	71.6	–	–	–	–	–	–	68.9	67.5	66.8	–
Science																			
GPQA diamond (single, pass@1)	**73.7**	**69.8**	**57.2**	49.0	68.4	84.0	64.7	60.1	51.5	71.4	53.6	78.2	68.0	80.2	71.5	**90.0**	42.4	46.0	79.7
GPQA diamond (multiple)	–	–	–	–	–	–	–	–	–	–	–	**84.6**	–	–	–	–	–	–	–
Mathematics																			
AIME 2025 (single, pass@1)	–	–	–	–	86.5	–	–	–	–	–	–	49.5	–	77.3	70.0	–	–	–	86.7
AIME 2025 (multiple)	–	–	–	–	–	–	–	–	–	–	–	–	**93.3**	–	–	–	–	–	–
AIME 2024 (single, pass@1)	–	–	–	–	87.3	–	–	–	–	36.7	–	61.3	–	83.9	79.8	–	–	–	92.0
AIME 2024 (multiple)	–	–	–	–	–	–	–	–	–	–	–	–	**93.3**	–	–	–	–	–	–
MATH-500	**95.0**	–	–	–	–	–	91.8	–	–	–	–	–	82.2	–	–	–	–	–	–
Code Generation / Coding																			
LiveCodeBench v5 (single, pass@1)	–	–	–	–	**74.1**	–	–	–	–	–	–	–	70.4	–	–	–	–	–	–
LiveCodeBench v5 (multiple)	–	–	–	–	–	–	–	–	–	–	–	–	**79.4**	–	–	–	–	–	–
LiveCodeBench (10/24–02/25)	**49.4**	**43.4**	32.8	27.7	**45.8 / 49.2**[c]	–	36.0	34.5	28.9	–	32.3	–	–	–	**33.3**	29.7	–	–	
Code Editing																			
Aider Polyglot	–	–	–	–	60.4[d]	74.0 / 68.6	–	–	–	44.9[d]	–	64.9[d]	–	–	56.9[d]	–	–	–	–
Agentic Coding																			
SWE-bench verified	–	–	–	–	49.3	63.8	–	–	–	38.0	–	**70.3**	–	–	49.2	–	–	–	–
Factuality																			
SimpleQA	–	–	–	–	13.8	**52.9**	–	–	–	62.5	–	–	–	43.6	30.1	–	–	–	–
Visual / Image Reasoning																			
MMMU (single)	**76.1**	73.4	69.4	–	No MM[e]	**81.7**	72.7	71.7	68.6	74.4	69.1	75.0	71.8	–	76.0	No MM[e]	64.9	62.8	No MM[e]
MMMU (multiple)	–	–	–	–	–	–	–	–	–	–	–	–	**78.0**	–	–	–	–	–	–
MathVista	–	73.7	70.7	–	–	–	73.1	57.6	–	–	–	–	70.7	No MM[e]	No MM[e]	–	67.6	68.9	–
Image Understanding																			
Vibe-Eval (Reka)	–	–	–	–	–	69.4	–	–	–	–	–	–	No MM[e]	No MM[e]	–	–	–	–	No MM[e]
ChartQA	–	**90.0**	88.8	No MM[e]	–	–	88.3	73.0	–	–	–	85.7	–	–	–	No MM[e]	76.3	**86.2**	–
DocVQA (test)	–	**94.4**	**94.4**	No MM[e]	–	–	–	91.2	–	–	–	92.8	–	–	–	–	90.4	**94.1**	–
Long Context																			
MRCR (128k avg)	–	–	–	–	–	**94.5**	–	61.4	–	–	64.0	–	–	–	–	–	–	–	–
MRCR (1M pointwise)	–	–	–	–	–	**83.1**	–	–	–	–	–	–	–	–	–	–	–	–	–
MTOB (half book)	–	**54.0 / 46.4**	**42.2 / 36.6**	128k[f]	128k[f]	–	–	48.4 / 39.8	42.3 / 35.1	–	–	128k[f]	–	–	128k[f]	128k[f]	128k[f]	128k[f]	–
MTOB (full book)	–	**50.8 / 46.7**	**39.7 / 36.3**	–	–	–	–	45.5 / 39.6	35.1 / 30.0	–	–	128k[f]	–	–	128k[f]	–	–	–	–
Multilingual Performance																			
Global MMLU (Lite)	–	–	–	–	–	**89.8**	–	–	–	–	–	–	–	–	–	–	–	–	–
Multilingual MMLU	–	**84.6**	–	–	–	–	–	–	–	–	81.5	–	–	–	–	–	–	–	–
Multilingual MMLU (OpenAI)	**85.8**	–	–	–	–	–	–	–	–	85.1	–	–	–	–	–	–	–	–	–

[a] $0.19/1Mtok (3:1 blended) estimated distributed inference cost (Llama 4 Maverick).

[b] Text problems only.

[c] DeepSeek v3.1 internal result (45.8) used as range unknown for LiveCodeBench (10/24–02/25).

[d] Diff performance (Aider Polyglot).

[e] No multimodal support reported/applicable. Abbreviated as 'No MM'.

[f] Context window limits reported result (typically 128k). Abbreviated as '128k'.

General Notes: Scores are self-reported by vendors unless otherwise specified. Bold text () **indicates the highest score *in the original source table* for that benchmark row, not necessarily the highest across this combined table. ‘pass@1‘: Single attempt evaluation. ‘Multiple‘: Evaluation using multiple attempts/voting. Gemini 2.5 Pro results used model ‘gemini-2.5-pro-exp-03-25‘ with default sampling (pass@1). Llama 4 results (Maverick, Scout) are 0-shot, temp=0, averaged for high-variance benchmarks. Llama 4 Behemoth results are current best internal runs (preview model). Cost estimates for non-Llama models sourced from Artificial Analysis. Non-Gemini/non-Llama results represent the highest self-reported scores found in the source documents. Model names abbreviated in headers for space (L4 = Llama 4, L3.1 = Llama 3.1). ‘–‘ indicates data not found in sources.**

Sources: Google Gemini (`https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#enhanced-reasoning`), Llama (`https://www.llama.com`), SAFE (`https://agi.safe.ai/`), Math Arena (`https://matharena.ai/`), LiveCodeBench (`https://livecodebench.github.io/`), Aider Leaderboard (`https://aider.chat/docs/leaderboards`).