FORGE

OFFICIAL RANKINGS

Every Frontier AI Model. Scored. Ranked. Compared.

May 2026 Edition

The Forge: AI Builders & Creators

facebook.com/groups/theforgeai

Data sourced from Scale AI SEAL, Artificial Analysis, OpenAI, Anthropic, Google DeepMind, DeepSeek, Moonshot AI, BenchLM, HuggingFace, and official model cards.
Updated May 2, 2026.

WHAT YOU NEED TO KNOW

Five frontier models launched in two weeks. That has never happened before. Claude Opus 4.7 (April 16), Kimi K2.6 (April 21), GPT-5.5 (April 23), DeepSeek V4 (April 24), and Grok 4.3 (April 30) all dropped within 14 days of each other. The competitive gap between the top labs has never been thinner.

1. No single model wins every coding benchmark.

Claude Opus 4.7 leads SWE-bench Pro (real GitHub bugs). GPT-5.5 leads Terminal-Bench (command-line agents). DeepSeek V4 Pro leads LiveCodeBench and Codeforces. Aggregate scores: all three within 2 points. S-tier.

2. GPT-5.5 owns terminal and agentic workflows.

82.7% on Terminal-Bench 2.0 (command-line agent tasks). 13.3 points ahead of Opus 4.7. For DevOps automation, pipeline runners, server setup, GPT-5.5 is untouchable.

3. DeepSeek V4 broke the price-quality curve.

V4 Pro scores 80.6% on SWE-bench Verified at $1.74/$3.48 per million tokens. V4 Flash gets you 79% at $0.14/$0.28. That's 56x cheaper than Opus 4.7 at 90% of the quality.

4. Reasoning has converged at the frontier.

GPQA Diamond: Claude 94.2%, GPT-5.4 Pro 94.4%, Gemini 3.1 Pro 94.3%. The top four models are within 0.8 percentage points. Reasoning is no longer a differentiator.

5. Open-weight models now match closed-source on coding.

MiniMax M2.5 hits 80.2% SWE-bench Verified. DeepSeek V4 Pro hits 80.6%. Both are open-weight. Both cost a fraction of proprietary APIs. The gap has closed.

METHODOLOGY

How We Score

Each model receives a tier grade (S through D) in four categories. We use real benchmark data from independent sources where available, and flag vendor-reported scores accordingly. A model needs verified scores in at least 3 of 4 categories to receive a composite ranking. Models with less coverage are marked Partial Data.

Tier	Score Range	What It Means
S	95-100	Best in class. Clear leader. No close competition.
A	85-94	Frontier competitive. Production ready for demanding workloads.
B	75-84	Strong. Good enough for most production use cases.
C	65-74	Capable but noticeable gaps vs frontier models.
D	Below 65	Not competitive at this level of comparison.

Scoring Categories (Each 25%)

Coding (25%): SWE-bench Pro (primary, contamination-resistant), SWE-bench Verified (secondary, contamination flagged), Terminal-Bench 2.0, LiveCodeBench, Codeforces rating.

Agentic (25%): GDPval-AA (real-world agentic Elo), MCP-Atlas (tool calling), OSWorld-Verified (computer use), TAU2-bench, BrowseComp.

Reasoning (25%): GPQA Diamond (PhD-level science), Humanity's Last Exam (HLE), ARC-AGI-2 (abstract reasoning), Artificial Analysis Intelligence Index.

Practical (25%): Context window, API latency/throughput, license terms, multimodal support, self-hosting viability.

Contamination Warning

SWE-bench Verified is confirmed contaminated. OpenAI audited it and found every frontier model could reproduce verbatim gold patches for certain tasks. OpenAI stopped reporting Verified scores. We still include Verified because it's widely cited, but SWE-bench Pro is our primary coding benchmark. All Verified scores carry a contamination flag in our tables.

Scaffolding Matters

The same model can score 10-15 points higher with optimized scaffolding vs a standardized harness. Claude Opus 4.5 scores 45.9% on SWE-bench Pro with SEAL standardized scaffolding, but 55.4% with Claude Code. We note the scaffold for every score. When comparing models, compare scores from the same scaffold type.

THE CODING WARS

Two models. Two very different wins. Claude Opus 4.7 leads on SWE-bench Pro (fixing real GitHub issues across multi-language codebases). GPT-5.5 leads on Terminal-Bench 2.0 (command-line agentic work). DeepSeek V4 Pro quietly matches Opus 4.6 on SWE-bench Verified at a fraction of the cost.

SWE-bench Pro Rankings (Primary)

The harder, contamination-resistant benchmark. Multi-language. Real engineering tasks.

Model	SWE-Pro	Tier	Open	Cost/1M*	Source
Claude Opus 4.7	64.3%	S	No	$10.00	Anthropic internal
GPT-5.5	58.6%	A	No	$11.25	OpenAI internal
GPT-5.4 (xHigh)	57.7%	A	No	$4.38	Vendor reported
DeepSeek V4 Pro (Max)	55.4%	A	Yes	$0.55*	DeepSeek internal
Muse Spark	55.0%	A	No	~$3.00	Meta internal
Gemini 3.1 Pro	54.2%	A	No	$4.50	Google internal
Claude Opus 4.6	51.9%	B+	No	$10.00	SEAL standardized
DeepSeek V4 Flash (Max)	~48%	B	Yes	$0.18	Estimated
Claude Opus 4.5	45.9%	B	No	$10.00	SEAL standardized
Claude Sonnet 4.5	43.6%	B	No	$6.00	SEAL standardized
Gemini 3 Pro	43.3%	B	No	$4.50	SEAL standardized
Claude Haiku 4.5	39.5%	C	No	$2.00	SEAL standardized

*Blended cost at 3:1 input:output ratio. DeepSeek V4 Pro uses promo pricing through May 31.

Terminal-Bench 2.0 + SWE-bench Verified

Model	Terminal-B	SWE-Verif*	Tier	Open
GPT-5.5	82.7%	~85%	S	No
GPT-5.4	75.1%	~80%	A	No
Claude Opus 4.7	69.4%	87.6%	A	No
DeepSeek V4 Pro (Max)	67.9%	80.6%	A	Yes
Claude Opus 4.6	65.4%	80.8%	B+	No
MiniMax M2.5	-	80.2%	B+	Yes
DeepSeek V4 Flash	-	79.0%	B	Yes
Qwen 3.6 Plus	-	78.8%	B	Yes
MiMo V2.5 Pro	-	78.0%	B	Yes
GLM-5	-	77.8%	B	Yes

Model	Terminal-B	SWE-Verif*	Tier	Open
Kimi K2.5	-	76.8%	**B**	Yes
Grok 4	-	73.5%	**C+**	No

*SWE-bench Verified is confirmed contaminated. Included for reference only.

FORGE PICKS: CODING

Best for fixing real bugs across codebases: Claude Opus 4.7

Best for terminal/DevOps agents: GPT-5.5

Best value for coding: DeepSeek V4 Pro (90% quality at 7x less cost)

Best open-weight for self-hosting: DeepSeek V4 Pro or MiniMax M2.5

THE AGENTIC SHOWDOWN

Agentic benchmarks are the most fragmented category. No single benchmark covers everything. GDPval-AA tests economically valuable real-world tasks. MCP-Atlas tests tool calling. OSWorld tests computer use. BrowseComp tests web research. Different models win on different axes. That's the point.

Model	GDPval Elo	Term-Bench	MCP-Atlas	BrowseComp	Tier
GPT-5.5 (xhigh)	1785	82.7%	-	84.4%	S
Claude Opus 4.7	1753	69.4%	77.3%	79.3%	A
Claude Opus 4.6	~1750	65.4%	75.8%	-	A
GPT-5.4	1674	75.1%	68.1%	89.3%	A
Grok 4.3	1500	-	-	-	B+
DeepSeek V4 Pro	-	67.9%	73.6%	-	B+
Gemini 3.1 Pro	1314	-	73.9%	85.9%	B
Kimi K2.6	-	-	-	-	B
Kimi K2 (K2.5)	-	-	-	-	B
Muse Spark	-	-	-	-	B

Grok 4.3 jumped 321 Elo points on GDPval-AA from its 4.20 version (1179 to 1500). That's the largest single-version agentic improvement we've tracked. Its multi-agent architecture (four parallel sub-agents debating in real-time) is genuinely novel.

Kimi K2.6 ranks #7 out of 115 models on BenchLM's Agentic category with a score of 87.9. Independent standardized scores are still arriving. As an open-weight model built specifically for agentic work, it's worth watching closely.

FORGE PICKS: AGENTIC

Best for terminal agents and DevOps: GPT-5.5 (untouchable on Terminal-Bench)

Best for tool calling and multi-step agents: Claude Opus 4.7 (leads MCP-Atlas)

Best for web research agents: GPT-5.4 Pro or Gemini 3.1 Pro (BrowseComp leaders)

Best open-weight agentic model: DeepSeek V4 Pro or Kimi K2.6

THE REASONING HEAVYWEIGHTS

Reasoning has converged at the frontier. The top four closed-source models are within 0.8 points on GPQA Diamond. The interesting story isn't who's #1. It's that reasoning is no longer the differentiator. The differentiation has moved to coding and agentic execution.

Model	GPQA-D	HLE (tools)	ARC-AGI-2	AA Index	Tier
GPT-5.5 (xhigh)	93.6%	52.2%	85.0%	60	S
Claude Opus 4.7 (max)	94.2%	54.7%	-	57	S
Gemini 3.1 Pro	94.3%	-	-	57	S
GPT-5.4 (xhigh)	94.4%	-	73.3%	57	S
Kimi K2.6	-	-	-	54	A
MiMo-V2.5-Pro	-	-	-	54	A
Grok 4.3	-	-	-	53	A
Claude Opus 4.6	91.3%	-	-	53	A
DeepSeek V4 Pro (Max)	90.1%	-	-	52	A
Muse Spark	-	-	-	52	A
Claude Sonnet 4.6	-	-	-	51	A

Arena Elo (LMSYS, March 2026): Provider Rankings

Provider	Arena Elo	Rank
Anthropic (Claude)	1,503	#1
xAI (Grok)	1,495	#2
Google (Gemini)	1,494	#3
OpenAI (GPT)	1,481	#4
Alibaba (Qwen)	1,449	#5
DeepSeek	1,424	#6

Top 4 providers separated by just 22 Elo points. The competitive gap is effectively closed.

HALLUCINATION WARNING

GPT-5.5 has the highest knowledge accuracy (57% on AA-Omniscience) but an 86% hallucination rate. Claude Opus 4.7 hallucinates at 36%. MiMo-V2.5-Pro at 25%. Grok 4.20 at 17%. For agents where being wrong is worse than not answering, this matters more than any reasoning score.

FORGE PICKS: REASONING

Best raw reasoning: GPT-5.5 (AA Index 60, but 86% hallucination rate)

Best reliable reasoning: Claude Opus 4.7 (AA Index 57, 36% hallucination, lowest at frontier)

Best value for reasoning: Gemini 3.1 Pro (AA Index 57 at $4.50 vs $11.25 for GPT-5.5)

THE PRICE WAR

This is the table that changes how you think about AI infrastructure. Per-token price is not per-task cost. Some models use 4x more tokens for the same work. The cost gap is 56x. But the effective gap depends on what you're building.

Model	In $/M	Out $/M	Blended	SWE-Verif	Context	Open
DeepSeek V4 Flash	$0.14	$0.28	$0.18	79.0%	1M	Yes
DeepSeek V3.2	$0.28	$0.42	$0.32	73.0%	128K	Yes
DeepSeek V4 Pro*	$0.44	$0.87	$0.55	80.6%	1M	Yes
Gemini 3 Flash	$0.50	$3.00	$1.13	78.0%	1M	No
Kimi K2.5	$0.60	$2.50	$1.08	76.8%	256K	Yes
MiMo-V2.5-Pro	~$0.50	~$2.50	$1.50	78.0%	128K	Yes
Grok 4.3	~$1.00	~$3.00	$1.60	-	1M	No
Kimi K2.6	$0.95	$4.00	$1.71	~80%	256K	Yes
Claude Haiku 4.5	$1.00	$5.00	$2.00	73.3%	200K	No
GPT-5.4	$2.50	$10.00	$4.38	~80%	1M	No
Gemini 3.1 Pro	$2.00	$12.00	$4.50	80.6%	2M	No
Claude Sonnet 4.6	$3.00	$15.00	$6.00	79.6%	200K	No
Claude Opus 4.7	$5.00	$25.00	$10.00	87.6%	1M	No
GPT-5.5	$5.00	$30.00	$11.25	88.7%	1M	No

*DeepSeek V4 Pro promotional pricing through May 31, 2026. Standard: $1.74/$3.48.

The 56x Cost Gap

Metric	Opus 4.7	DS V4 Flash	Gap
Input per 1M tokens	$5.00	$0.14	36x
Output per 1M tokens	$25.00	$0.28	89x
Blended per 1M tokens	$10.00	$0.18	56x
100M tokens/day (monthly)	$30,000	$540	56x
SWE-bench Verified	87.6%	79.0%	8.6pts

FORGE PICKS: PRICE

Cheapest production-viable model: DeepSeek V4 Flash ($0.18 blended)

Best quality-per-dollar: DeepSeek V4 Pro at promo pricing ($0.55 blended, 80.6% SWE-Verified)

Best mid-tier value: Gemini 3.1 Pro ($4.50 blended, frontier reasoning at half Opus cost)

When to pay premium: Opus 4.7 for mission-critical coding, GPT-5.5 for terminal agents

THE SELF-HOSTER'S CORNER

For builders running models on their own hardware, API pricing is just one option. Self-hosting brings the per-token cost to near zero (just electricity and hardware amortization). Here's every open-weight model worth running, ranked by practical self-host viability.

Model	Parameters	VRAM (est.)	License	SWE-Verif	Tier
DeepSeek V4 Pro	1.6T MoE (49B active)	Multi-GPU cluster	MIT	80.6%	A
DeepSeek V4 Flash	284B MoE (13B active)	~40GB+ Q4	MIT	79.0%	A
Kimi K2.6	1T MoE	Multi-GPU	Modified MIT	~80%	A
MiniMax M2.5	-	-	-	80.2%	B+
Kimi K2.5	1T MoE (32B active)	~24GB Q4 active	Modified MIT	76.8%	B+
Qwen 3.6 Plus	-	-	Apache 2.0	78.8%	B+
GLM-5.1	744B MoE (40B active)	~30GB Q4 active	MIT	77.8%*	B+
MiMo V2.5 Pro	1T	Large	-	78.0%	B
Gemma 4 31B	31B dense	~20GB Q4	Apache 2.0	-	B
Qwen 3.5-122B-A10B	122B MoE (10B active)	~12GB Q4 active	Apache 2.0	-	B
Llama 4 Maverick	400B MoE (17B active)	~15GB Q4 active	Meta License	-	C+

*GLM-5/5.1 built on Huawei Ascend chips without NVIDIA hardware.

Hardware Reference Points

Single RTX PRO 6000 (96GB VRAM): Can run Qwen 3.5-122B-A10B, Gemma 4 31B, Llama 4 Maverick (active params), and smaller models comfortably at Q4 quantization. DeepSeek V4 Flash may fit with aggressive quantization.

Single RTX 4090 (24GB VRAM): Qwen 3.5-122B-A10B at Q4 (10B active, ~12GB), Gemma 4 27B at Q4 (~18GB tight), smaller Qwen/Gemma variants.

Multi-GPU cluster (2-8x 80GB+): Required for full DeepSeek V4 Pro, Kimi K2.6, and other trillion-parameter MoE models. Even with FP4 quantization, these models need serious hardware.

FORGE PICKS: SELF-HOSTING

Best single-GPU model: Qwen 3.5-122B-A10B (only 10B active, fits on a 4090)

Best if you have serious hardware: DeepSeek V4 Flash (near-frontier at 284B MoE)

Best license for commercial use: Gemma 4 or Qwen (Apache 2.0, zero restrictions)

THE FORGE PICKS

Opinionated recommendations from someone who runs multi-agent systems in production, self-hosts on real hardware, and actually pays these API bills. Your mileage may vary. But this is where I'd put my money in May 2026.

"I'm building a coding agent that fixes production bugs."
Claude Opus 4.7. Nothing else comes close on SWE-bench Pro. If budget matters, pair Sonnet 4.6 for routine tasks and route hard problems to Opus.

"I'm building DevOps / terminal automation agents."
GPT-5.5. The Terminal-Bench lead is 13 points. For shell scripting, server provisioning, CI/CD pipelines, it's not a contest.

"I need the smartest model available, cost be damned."
GPT-5.5 (xhigh) for broadest capability. Claude Opus 4.7 (max) if coding is the priority. They're functionally tied on reasoning. Pick by task.

"I want frontier-level coding at the lowest possible cost."
DeepSeek V4 Pro at promotional pricing. 80.6% SWE-bench Verified at $0.55 blended. That's 18x cheaper than Opus 4.7 at 92% of the quality.

"I need to process millions of tokens cheaply."
DeepSeek V4 Flash. $0.14 per million input tokens. Nothing else is in the same universe at this price point. Use it for classification, extraction, and routing.

"I want to self-host on my own GPU."
Qwen 3.5-122B-A10B if you've got a single GPU. DeepSeek V4 Flash if you've got more headroom. Both under Apache 2.0 / MIT with no commercial restrictions.

"I need the best all-around mid-tier model."
Gemini 3.1 Pro. Frontier reasoning at $4.50 blended. 2M context window. Half the cost of Opus, competitive on almost everything except coding.

"I'm building something with real-time web research."
GPT-5.4 Pro (BrowseComp 89.3%) or Gemini 3.1 Pro (85.9%). Opus 4.7 regressed on BrowseComp. Don't use it for web research agents.

WHAT TO WATCH: JUNE 2026

Five models dropped in two weeks. The dust hasn't settled yet. Here's what we're tracking for the June edition.

Independent benchmarks for GPT-5.5 and Grok 4.3. Both are under two weeks old. The SEAL standardized scores, Artificial Analysis full suite, and community testing will reshape some of these rankings. Expect Grok 4.3 to climb once independent data arrives.

DeepSeek V4 Pro post-promotional pricing. The current $0.44/$0.87 promo expires May 31. Standard pricing jumps to $1.74/$3.48. Still cheap, but the value math shifts. We'll rescore accordingly.

Kimi K2.6 full coverage. Moonshot's latest dropped April 21 but independent benchmarks are thin. BenchLM has it at #7 in agentic work. We need SWE-bench Pro, GPQA Diamond, and GDPval scores before we can give it a full composite rating.

Claude Mythos public availability. Currently restricted to 50 organizations under Project Glasswing. If Anthropic opens access, it rewrites every table in this report. 93.9% SWE-bench Verified. 94.6% GPQA Diamond. We'll be watching.

ABOUT THE FORGE

The Forge: AI Builders & Creators is a community for developers, entrepreneurs, and builders who ship real products with AI. Not hype. Not theory. Actual implementation. We share model evaluations, production architectures, self-hosting configurations, and the tools that make multi-agent systems work in the real world.

This report is published monthly and distributed free to group members. Every score is sourced from independent benchmarks or official model cards. Every recommendation comes from production experience, not press releases. If you're building with AI and want signal instead of noise, you're in the right place.

The Forge: AI Builders & Creators

facebook.com/groups/theforgeai

Updated monthly. Next edition: June 2026.

Data sourced from Scale AI SEAL, Artificial Analysis, official model cards, and HuggingFace.
Benchmark scores are point-in-time and may change as independent evaluations arrive.
This report is for informational purposes only. Always verify current pricing with providers.