AIVERIS

AIVERIS REPORT · APRIL 23, 2026

State of the Art

Frontier Large Language Models

A comprehensive comparative analysis by Aiveris, LLC

Prepared for decision-makers evaluating LLM procurement, routing, and deployment strategy.

aiveris.ai

CONTENTS

01

Executive Summary

In the ten weeks since Aiveris' February 16, 2026 analysis, the frontier has moved sharply but fragmented further. No single model dominates every dimension, and the competitive gap between the top three labs has compressed to statistical noise on composite benchmarks while widening significantly on specific capability axes.

Headline shifts since February 16, 2026

- **Anthropic Claude Opus 4.7** (released April 16, 2026) now leads software engineering: 87.6% on SWE-bench Verified and 64.3% on SWE-bench Pro (vendor-reported), displacing Opus 4.6 as the publicly-available coding leader. Anthropic has also disclosed an unreleased internal model, Claude Mythos Preview, used only within "Project Glasswing" — it is not generally available and is disclosed as stronger than Opus 4.7.
- **OpenAI GPT-5.4** shipped March 5, 2026 (unifying the Codex and mainline tracks), and **GPT-5.5** (codenamed "Spud") was released April 23, 2026, to paid ChatGPT subscribers with API access following. GPT-5.5 claims state-of-the-art 82.7% on Terminal-Bench 2.0 and 58.6% on SWE-bench Pro. API pricing is higher than GPT-5.4 ($5/$30 per MTok standard, with a GPT-5.5-pro variant at $30/$180).
- **Google Gemini 3.1 Pro** (preview since February 19, 2026) and an upgraded Gemini 3 Deep Think now lead PhD-level reasoning: 94.1% GPQA Diamond and 48.4% on Humanity's Last Exam (HLE) without tools for Deep Think.
- **Meta Muse Spark** (April 8, 2026) is Meta's first frontier-competitive release, and a strategic break from open weights — it is proprietary. It ranks fourth at Artificial Analysis Intelligence Index 52.
- **Moonshot Kimi K2.6** (April 18–20, 2026) refreshes the open-weight crown from K2.5: 1T-parameter MoE, 256K context, 54.0% HLE w/ tools, 58.6% SWE-bench Pro.
- **Zhipu GLM-5.1** (open weights, April 7, 2026 under MIT) and **Alibaba Qwen 3.6-Plus / Qwen 3.6-35B-A3B** (April 2026) narrow the open-closed gap to a handful of points on coding benchmarks.
- **xAI Grok 4.20** (multiple beta rollouts through April) and **Grok 4.3 Beta** (April 17, 2026, SuperGrok Heavy tier) retain a 2M-token context window and industry-leading inference speed but score below the top three on composite benchmarks.

Category leadership as of April 23, 2026

DOMAIN	LEADER	RUNNER-UP
Agentic software engineering (SWE-bench Pro)	Claude Opus 4.7 — 64.3% (Anthropic-reported); GPT-5.5 — 58.6% (OpenAI-reported)	GLM-5.1 — 58.4% (open weight)
Agentic software engineering (SWE-bench Verified)	Claude Opus 4.7 — 87.6%	GPT-5.3-Codex — 85.0%
Terminal-Bench 2.0	GPT-5.5 — 82.7% (vendor), with ForgeCode + Opus 4.6 / GPT-5.4 tied at 81.8% on the public board	TongAgents + Gemini 3.1 Pro — 80.2%
Human preference (LMArena Text)	Claude Opus 4.7 / Claude Opus 4.6 Thinking — ~1504 Elo	Gemini 3.1 Pro Preview — ~1493 Elo
Graduate-level science (GPQA Diamond)	Gemini 3.1 Pro — 94.1%	GPT-5.4 — 92.0% / GPT-5.3-Codex — 91.5%
Humanity's Last Exam (HLE)	Gemini 3 Deep Think — 48.4% (no tools)	GPT-5.4 xhigh — 41.6%; Kimi K2.6 — 54.0% w/ tools
ARC-AGI-2	Gemini 3 Deep Think — 84.6%	Gemini 3.1 Pro — ~77.1% (vendor)
ARC-AGI-3 (new, March 27, 2026 launch)	All frontier models below 1% — benchmark wide open	—
Frontier mathematics (AIME 2025)	Saturated: GPT-5.4, Gemini 3 Pro, Deep Think all at 98–100% (tools)	—
Open-weight overall	Kimi K2.6 — AA Intelligence Index 54	GLM-5.1 (Intelligence Index 51)
Cost-efficiency at frontier-adjacent quality	DeepSeek V3.2 — $0.28/$0.42 per MTok	Qwen 3.6-Plus (free preview) / GLM-5.1 ($1.00/$3.20)
Multimodal + long context	Gemini 3.1 Pro — 1M tokens, native A/V	Grok 4.20 / 4.3 — 2M tokens (largest Western closed context)

Key strategic conclusion: The "one model to rule them all" thesis is dead for 2026. Anthropic owns coding; Google owns reasoning and multimodality; OpenAI owns the agentic-terminal and GDPval economic-value axis; the Chinese open-weight bloc (Moonshot, Zhipu, Alibaba, DeepSeek) now undercuts Western leaders by 5–50× on cost while trailing by 3–10 points on the hardest benchmarks. Routing — not procurement — is the binding constraint on enterprise LLM economics.

02

Benchmark Methodology and Taxonomy

Benchmarks used in this report fall into six evaluative families; each family has a different failure mode that Section 7 re-visits explicitly.

1. **Software engineering:** SWE-bench Verified (500 Python tasks, partly contaminated), SWE-bench Pro (1,865 tasks across Python/Go/TypeScript/JavaScript, contamination-resistant), SWE-rebench (Nebius' rolling-window benchmark sourced after model release dates), SWE-bench Live (monthly refreshed), Aider Polyglot (multi-language editing).
2. **Agentic automation:** Terminal-Bench 2.0 (the current version at tbench.ai; Terminal-Bench 3.0 is in development per the registry), OSWorld, Vending-Bench 2, APEX-Agents, GDPval/GDPval-AA.
3. **Human preference:** LMArena / Arena Intelligence's Bradley-Terry Elo system across Text, Code, Vision, Search, Document, WebDev, Image, and Video sub-arenas (rebranded from LMArena to "Arena" on January 28, 2026).
4. **Academic and scientific reasoning:** MMLU-Pro, MMMU / MMMU-Pro, GPQA Diamond, MMMLU (multilingual).
5. **Frontier-hard:** Humanity's Last Exam (HLE), ARC-AGI-1 (approaching saturation), ARC-AGI-2, ARC-AGI-3 (released March 27, 2026), FrontierMath, MathArena Apex.
6. **Mathematical competition:** AIME 2025 (increasingly saturated), AIME 2026 (I/II), USAMO 2026, International Mathematical Olympiad and International Olympiad in Informatics gold-medal evaluations, HMMT.

Composite indices considered: Artificial Analysis Intelligence Index v4.0.4 (10 evaluations across four equally-weighted categories: Agents, Coding, General, Scientific Reasoning) and Vellum / BenchLM composite scores. We flag composite scores as directional only; component-level reading (Section 6) is always preferred.

Throughout, we differentiate three tiers of evidence:

- Vendor-reported scores (noted explicitly; flagged skeptically).
- Independent evaluation (Artificial Analysis, Vals.ai, Scale AI SEAL, Nebius SWE-rebench, Epoch AI, Vellum).
- Live human preference (Arena).

03

Frontier Model Profiles

3.1 Anthropic

Claude Opus 4.7

Flagship as of April 16, 2026. Anthropic's most capable generally available model. Released April 16, 2026, replacing Opus 4.6 in the flagship slot.

- **Pricing:** $5 per million input tokens / $25 per million output tokens — unchanged from Opus 4.6. Prompt caching up to 90% discount; 50% batch discount; 1.1× surcharge for US-only inference.
- **Tokenizer caveat:** A new tokenizer can increase token counts by 1.0×–1.35× relative to Opus 4.6 on the same content, concentrated on code, structured data, and non-English text — effective cost can rise up to 35% despite unchanged per-token pricing.
- **Context:** 1M tokens, included at standard pricing; max output 128K (300K via output-300k-2026-03-24 batch beta).
- **New controls:** "xhigh" effort level between high and max; new "task budgets"; Claude Design product; Claude Code review tools.
- **Benchmark performance** (Anthropic-reported unless otherwise noted): SWE-bench Pro 64.3% (up from Opus 4.6's 53.4–51.9%); SWE-bench Verified 87.6%; Terminal-Bench 2.0 69.4% (not yet on tbench.ai public board); CursorBench 70%; LMArena Text Elo ~1504 (independent).
- **Strengths:** Agentic coding, long-horizon autonomy (Devin reports hour-scale coherent runs), high-resolution vision (tripled from Opus 4.6 to 3.75 MP), professional enterprise workflows (slides, docs, spreadsheets).
- **Limitations:** High token cost per task; Anthropic publicly concedes Opus 4.7 does not match their unreleased Mythos Preview; first Anthropic model rated High cyber capability under Preparedness Framework — deployed with automated safeguards that block identified high-risk cybersecurity requests.

Claude Opus 4.6

Still available; $5/$25 per MTok; 1M context; SWE-bench Verified 80.8%, SWE-bench Pro 51.9% on Scale SEAL's mini-swe-agent; ForgeCode + Opus 4.6 tied at 81.8% on Terminal-Bench 2.0 public board; Code Arena Elo 1542–1548.

Claude Sonnet 4.6

Released February 17, 2026. $3/$15 per MTok; 1M context at standard pricing; SWE-bench Verified 79.6%. The default model for Free and Pro users on claude.ai; 40% cheaper than Opus per token.

Claude Haiku 4.5

$1/$5 per MTok; 200K context; 64K max output. First Haiku with extended thinking, computer use, and context awareness; SWE-bench Verified 73.3%. The strongest Haiku-class cost/speed tier for sub-agents and high-volume pipelines.

Claude Mythos Preview

Not generally available. Released only to "Project Glasswing" cybersecurity consortium in early April 2026. Anthropic reports 93.9% SWE-bench Verified, 77.8% SWE-bench Pro, and 64.7% HLE on leaderboard tracker LLM-Stats. Because Mythos is a restricted-access consortium model and not purchasable or procurable by general enterprises, it is excluded from routing recommendations below — but it establishes an upper bound on publicly-auditable progress.

3.2 OpenAI

GPT-5.5 ("Spud")

Flagship as of April 23, 2026. Released April 23, 2026, the same day this report is dated.

- **Pricing** (effective when API launches): $5 per million input tokens, $30 per million output tokens; batch/flex at half rate; priority at 2.5×.
- **GPT-5.5-pro variant:** $30 / $180 per MTok.
- **Context:** 1M tokens (API); 400K tokens in Codex. Max output 128K.
- **Availability at launch:** ChatGPT Plus/Pro/Business/Enterprise and Codex; API "very soon" pending cybersecurity guardrails.
- **Vendor-reported benchmarks:** Terminal-Bench 2.0 82.7% (state-of-the-art claim), SWE-bench Pro 58.6%, better-than-GPT-5.4 on Expert-SWE (20-hour-median long-horizon internal eval).
- **Strengths:** Token-efficient on Codex workloads relative to GPT-5.4; stronger agentic autonomy ("hold context across large systems, reason through ambiguous failures").
- **Limitations:** Independent evaluation not yet available (less than 24 hours old at publication); "state-of-the-art" claim on Terminal-Bench 2.0 is vendor-reported and has not yet been confirmed on the public tbench.ai leaderboard. Pricing is higher than GPT-5.4 for the same nominal capability class.

GPT-5.4

Flagship March 5 – April 23, 2026. Unified successor to GPT-5.3-Codex and GPT-5.2; first OpenAI mainline model with native computer use, five-level reasoning effort control (none / low / medium / high / xhigh).

- **Pricing:** $2.50 / $15 per MTok for <272K input; $5.00 / $22.50 per MTok above 272K; GPT-5.4-pro $30 / $180 per MTok.
- **Context:** 272K standard, 1M (opt-in) in API and Codex; 128K max output.
- **Artificial Analysis Intelligence Index:** 57 (xhigh) — tied with Claude Opus 4.7 and Gemini 3.1 Pro Preview for first place.
- **Vendor benchmarks:** SWE-bench Pro 59.1% (Scale SEAL mini-swe-agent with xhigh), OSWorld 75% (first model to exceed human-expert baseline on desktop automation), ARC-AGI ~93.7%, HLE xhigh 41.6%, GPQA 92.0%, AIME 2025 100% with code execution.
- **Strengths:** Computer-use is the state of the art, and the $2.50/$15 standard pricing undercuts Claude Opus by 2× on input and 1.67× on output at similar composite intelligence.
- **Limitations:** Above-272K surcharge doubles input cost; GPT-5.2 is scheduled to retire June 5, 2026, forcing migration; Thinking message cap on Plus plan (80 per 3-hour window) bites heavy users.

GPT-5.3-Codex

Released February 5, 2026. First model OpenAI deployed as High capability under the Preparedness Framework for cybersecurity. SWE-bench Verified 85.0%, SWE-bench Pro 56.8%, Terminal-Bench 2.0 77.3%. Remains selectable in Codex for coding-specialized workflows.

GPT-5.3-Codex-Spark

Feb 12, 2026 research preview; first OpenAI model designed for real-time coding on Cerebras WSE-3, delivering >1000 tokens per second on 128K context. Not priced in the API yet; ChatGPT Pro-only preview.

GPT-5.4-mini

$0.40/$1.60 per MTok; ~94% of GPT-5.4 Standard's coding performance. The cost-efficient default for lower-complexity coding agent sub-tasks.

3.3 Google DeepMind

Gemini 3.1 Pro

Preview, flagship since February 19, 2026. Builds on Gemini 3 Pro (November 2025) with improved agentic reliability, stronger SWE performance, and more efficient token use.

- **Pricing:** $2 per million input tokens / $12 per million output tokens for ≤200K input; $4/$18 above 200K.
- **Context:** 1M tokens input, 64K output. Knowledge cutoff January 2025.
- **Artificial Analysis Intelligence Index:** 57 — tied at the top.
- **Benchmarks** (mix of vendor and independent): GPQA Diamond 94.1% (#1), HLE 44.7% (#1 non-tool-augmented — the single-strongest frontier HLE score), SWE-bench Verified 80.6%, SWE-bench Pro 46.1% on Scale SEAL mini-swe-agent, Terminal-Bench 2.0 up to 80.2% (TongAgents scaffold), ARC-AGI-2 77.1%, MMMU-Pro 81.0%, Video-MMMU 87.6%, Vending-Bench 2 top.
- **Strengths:** Native multimodal input (text/image/audio/video) in a single model; best single-shot HLE score; most competitive price-per-intelligence in the closed-weight tier.
- **Limitations:** Preview status (API behavior can change); single-thread latency is higher than GPT-5.4 at long context.

Gemini 3 Deep Think

Updated February 12, 2026. Specialized reasoning mode for Google AI Ultra subscribers, now available via Gemini API to select researchers/enterprises.

- **Benchmarks:** HLE 48.4% (no tools) — the highest tool-free HLE score on record; ARC-AGI-2 84.6% (verified by ARC Prize Foundation); Codeforces Elo 3455; gold-medal-level on IMO 2025 and IPhO/IChO 2025 written sections; CMT-Benchmark 50.5%; ARC-AGI-1 96%+ at $7.17/task.
- **Pricing:** Not publicly listed as a separate SKU; bundled with Gemini AI Ultra subscription, with API preview gated.
- **Strengths:** Best-in-class on agentic science/engineering and on benchmarks ARC Prize Foundation independently verifies.
- **Limitations:** Latency per task is high; pricing opacity and per-task cost on ARC-AGI-2 ($13.62/task) makes production routing decisions harder than with metered APIs.

Gemini 3 Pro, Gemini 3 Flash, Gemini 3.1 Flash-Lite

All in the 3-family. Flash-Lite at $0.25/M input is the cheapest tier at Google, and is the most-used Gemini SKU on free-tier preview. Context window consistent at 1M.

3.4 xAI

Grok 4.3 Beta

Flagship as of April 17, 2026. SuperGrok Heavy tier only ($300/month). Native video input, downloadable PDF/XLSX/PPTX outputs, tighter "Grok Computer" integration.

- **Architecture** (reported, not officially disclosed): ~0.5T params at launch with a 1T checkpoint in training; retains 16-agent Heavy system; 2M-token context window — the largest among Western closed-weight models.
- **Limitations noted:** No persistent memory between sessions; no full model card.

Grok 4.20

Current API flagship. $2/$6 per MTok; 2M context; Artificial Analysis Intelligence Index 49; output speed 173 t/s (notably fast for a reasoning model); TTFT 12.58s (high). LMArena Text Elo ~1491. xAI released Grok TTS ($4.20/1M chars) and Grok STT APIs in March/April 2026 — meaningfully cheaper than ElevenLabs/OpenAI for speech workloads.

3.5 Meta

Muse Spark

April 8, 2026. Meta Superintelligence Labs' first frontier model, and Meta's first proprietary flagship since the Llama era. Natively multimodal with "Contemplating Mode" multi-agent orchestration.

- **Availability:** Meta AI app, meta.ai, and a private API preview.
- **Artificial Analysis Intelligence Index:** 52 (fourth globally behind Opus 4.7, Gemini 3.1 Pro, GPT-5.4).
- **Vendor benchmarks:** SWE-bench Verified 77.4%, SWE-bench Pro 55.0% (SEAL), HLE 58% in Contemplating Mode, FrontierScience Research 38%. Token efficiency claim: 58M tokens on Intelligence Index vs 157M for Opus 4.6 and 120M for GPT-5.4.
- **Strengths:** Multimodal vision; health-domain reasoning (trained with >1,000 physicians); token efficiency.
- **Limitations:** Meta explicitly concedes "current performance gaps" in long-horizon agentic systems and coding workflows. On Terminal-Bench Hard, Muse Spark trails both Claude Sonnet 4.6 and GPT-5.4; on GDPval-AA, Muse Spark scores 1,427 Elo vs Claude Sonnet 4.6 at 1,648. No open weights; no API-of-record pricing published as of April 23, 2026.

04

The Open-Weight Ecosystem: Structural Shift Analysis

The February 2026 Aiveris report characterized open-weight Chinese models as "2–4 points behind the frontier on coding at 1/50th the cost." That framing is now out of date in three important ways.

(A) The gap is now <1 point on the single most-watched frontier benchmark. On April 7, 2026, Zhipu released GLM-5.1 under MIT license with a vendor-reported 58.4% on SWE-bench Pro — ahead of both GPT-5.4 (57.7%) and Claude Opus 4.6 (57.3%), and 5.9 points behind Claude Opus 4.7. GLM-5 (744B MoE, 40B active, DeepSeek Sparse Attention, 200K context, trained on 100K Huawei Ascend 910B chips) and GLM-5.1 (same architecture, RL-refined) are the first open-weight models explicitly trained without Nvidia hardware.

(B) Meta has exited open weights at the frontier. Muse Spark is proprietary; Meta's blog mentions "hope to open-source future versions" but commits nothing. This vacates the Western-lab frontier-open-weights slot that Llama occupied through 2024–2025.

(C) Google has filled that vacuum with Gemma 4 (April 2, 2026). Four variants (E2B, E4B, 26B-A4B MoE, 31B dense), Apache 2.0 license, multimodal on E2B/E4B including phone-grade deployments. Gemma 4 is not frontier-competitive with Opus 4.7 / GPT-5.4 / Gemini 3.1 Pro on hardest benchmarks, but its on-device performance meaningfully pushes the floor of what runs on laptops and phones.

Current open-weight tier (April 23, 2026)

MODEL	PARAMS (ACTIVE)	CONTEXT	PRICING (HOSTED)	COMPOSITE RANK
Moonshot Kimi K2.6	~1T / ~32B MoE	262K	$0.95 in / $4 out (Cloudflare Workers AI)	AA Intelligence Index 54 — #1 open weight (Artificial Analysis)
Moonshot Kimi K2.5	1T / 32B	256K	$0.60/M in	Earlier flagship; still widely deployed
Zhipu GLM-5.1	744B / 40B	200K	$1.00/$3.20 (Z.ai) or free weights	AA IIndex 51; SWE-bench Pro #1-tier open weight
Zhipu GLM-5	744B / 40B	200K	Similar	AA IIndex 50

MODEL	PARAMS (ACTIVE)	CONTEXT	PRICING (HOSTED)	COMPOSITE RANK
Alibaba Qwen 3.6-Plus (Preview)	Dense+MoE family	—	Free via OpenRouter preview	Entered Arena Code top 7 (Arena changelog)
Alibaba Qwen 3.6-35B-A3B	35B / ~3B MoE	—	Self-host	Efficient on-device tier
Alibaba Qwen3-Coder-Next	80B / 3B MoE	—	Self-host	70.6% SWE-bench Verified at 3B active
DeepSeek V3.2 (Non-reasoning)	685B / 37B MoE, MIT	128K	$0.28 / $0.42 per MTok (DeepSeek API)	AA IIndex 32; cheapest frontier-adjacent option (Artificial Analysis)
DeepSeek V3.2-Speciale	Same family	164K	$0.40 / $1.20	Stronger thinking+tool mode
MiniMax M2.5 / M2.7	Open weight	—	Hosted	SWE-bench Verified 80.2% (#1 open on some boards)
MiMo-V2-Pro (Xiaomi)	—	—	—	SWE-bench Verified 78.0%

DeepSeek V4 is still unreleased as of April 23, 2026, despite repeated rumors. DeepSeek remains the cheapest widely-accessible frontier-adjacent tier, but has fallen behind GLM-5.1 and Kimi K2.6 on composite rankings. Llama 4 Scout retains the largest open-weight context at 10M tokens.

The strategic shift: open-weight leadership has consolidated in four Chinese labs (Moonshot, Zhipu, Alibaba, DeepSeek), with Google Gemma 4 as the primary Western open-weight alternative. Enterprise open-weight procurement now entails more geopolitical review than it did in February.

05

Comparative Rankings by Capability Domain

5.1 Software Engineering

SWE-bench Verified (500 Python tasks, human-validated but partly contaminated per OpenAI's audit):

RANK	MODEL	SCORE	SOURCE
1	Claude Mythos Preview (restricted)	93.9%	llm-stats
2	Claude Opus 4.7	87.6%	Anthropic
3	GPT-5.3-Codex	85.0%	OpenAI
4	Claude Opus 4.5	80.9%	Anthropic
5	Claude Opus 4.6	80.8%	Anthropic
6	Gemini 3.1 Pro	80.6%	Google
7	MiniMax M2.5	80.2% (top open weight)	MiniMax
8	GPT-5.2	80.0%	OpenAI
9	Claude Sonnet 4.6	79.6%	Anthropic
10	Qwen 3.6-Plus	78.8%	Alibaba

Note: OpenAI's published Verified audit notes that 59.4% of the hardest unsolved Verified problems have flawed test cases, and that every major frontier model can reproduce some verbatim gold patches. Treat these as directional.

SWE-bench Pro (1,865 tasks across Python/Go/TS/JS, contamination-resistant by copyleft licensing and proprietary private set):

RANK	MODEL	PUBLIC SET %	SOURCE
1	Claude Mythos Preview (restricted)	77.8%	BenchLM, llm-stats / Anthropic
2	Claude Opus 4.7	64.3%	Anthropic
3	GPT-5.5	58.6%	

RANK	MODEL	PUBLIC SET %	SOURCE
			OpenAI (April 23, 2026, vendor)
4	Kimi K2.6	58.6%	Moonshot
5	GLM-5.1	58.4%	Zhipu (open weight)
6	GPT-5.4 xHigh	59.1% on SEAL mini-swe-agent	Scale SEAL
7	GPT-5.3-Codex (agent system)	56.8%	OpenAI
8	GPT-5.2-Codex	56.4%	OpenAI
9	Muse Spark	55.0%	Meta
10	Claude Opus 4.6 (SEAL mini-swe-agent)	51.9%	Scale SEAL

SWE-bench Pro public-private delta remains meaningful. Earlier frontier models dropped 5–8 points from public to private (e.g., GPT-5 from 23.1% public to 14.9% private in the September 2025 snapshot). When evaluating, bias toward Pro > Verified.

SWE-rebench (rolling 2026 window, contamination-controlled by post-release task sourcing):

RANK	MODEL	RESOLVED RATE
1	Claude Opus 4.6	65.3%
2	GLM-5 (open weight)	62.8%
3	DeepSeek V3.2	60.9%
4	Claude Sonnet 4.6	—
5	Kimi K2.5	—

Source: BenchLM SWE-rebench, April 10, 2026.

Approximate cost per successful SWE-bench-class fix (estimates using published pricing and token usage):

- Opus 4.7: ~$0.71/fix
- GPT-5.3-Codex: ~$1.06/fix (higher output tokens)
- Gemini 3.1 Pro: ~$0.93/fix
- Qwen 3.6-Plus: ~$0.25/fix

- DeepSeek V3.2: ~$0.23/fix

These figures are directional; actual cost per production fix is dominated by scaffolding overhead and verifier cost, not model billing.

5.2 Agentic Terminal Automation (Terminal-Bench 2.0)

Terminal-Bench 2.0 remains the de facto standard; Terminal-Bench 3.0 is in development but not yet public (tbench.ai).

RANK	MODEL + SCAFFOLD	SCORE
1	Claude Mythos Preview (restricted)	82.0% (llm-stats TB2)
2	GPT-5.5 (OpenAI-reported, pending board)	82.7%
3	ForgeCode + Claude Opus 4.6 / GPT-5.4 (tied)	81.8%
4	TongAgents + Gemini 3.1 Pro	80.2%
5	ForgeCode + Gemini 3.1 Pro	78.4%
6	SageAgent + GPT-5.3-Codex	78.4%
7	Claude Opus 4.7 (Anthropic-reported, pending board)	69.4%
8	GPT-5.3-Codex	77.3%
9	GPT-5.4	75.1%

Scaffold choice drives ±5–10 points; trust the combined (scaffold + model) row, not the raw model name.

5.3 Human Preference (LMArena / Arena Intelligence)

Arena rebranded from LMArena to "Arena" on January 28, 2026.

Text Arena top models (as of April 6–21, 2026):

RANK	MODEL	ELO
1	claude-opus-4-6-thinking	~1504
1	claude-opus-4-7 (just added April 17–20)	~1504 (approx, shares top band)
2	gemini-3.1-pro-preview	~1493

RANK	MODEL	ELO
3	gpt-5.4-high	~1482–1484
4	grok-4.20-beta-0309-reasoning	~1491
5	claude-sonnet-4-6	high 1480s

Source: Messenger Bot Chatbot Arena analysis, Ofox leaderboard, AI Dev Day. The top six are within a 24-Elo band — statistically meaningful but practically a coin-flip for general tasks.

Code Arena (April 9, 2026; 231,158 votes, 60 models):

- Claude Opus 4.6 Thinking: 1548 (+11/-11)
- Claude Opus 4.6: 1542 (+10/-10)
- Claude Sonnet 4.6: 1523
- GLM-5.1: 1530 (+20/-20) — strongest open-weight
- Claude Opus 4.5: 1465
- claude-opus-4-7-thinking added April 21, 2026

Vision Arena: Gemini 3 / 3.1 Pro top the closed board; claude-opus-4-7 and claude-opus-4-6-thinking added April 20.

Search Arena: Google and OpenAI now lead; xAI's Grok search advantage has eroded. gemini-3-flash-grounding, claude-opus-4-5-search, and gpt-5.2-search-non-reasoning were added over February–April.

WebDev Arena (Image-to-WebDev went live April 15): Claude Opus 4.7 leads on React multi-file.

5.4 Academic Reasoning and Scientific Knowledge

GPQA Diamond (Price Per Token leaderboard, April 16, 2026, 263 models):

- Gemini 3.1 Pro Preview: 94.1% (#1)
- GPT-5.4: 92.0% (#2)
- GPT-5.3-Codex: 91.5% (#3)
- Gemini 3 Pro: 91.9%
- Gemini 3 Deep Think: ~97% (internal)
- Claude Opus 4.6: ~91.3%

The 94.1% ceiling means GPQA Diamond is saturating in 2026; the spread of the top five fits inside a 3-point band. We now recommend discounting GPQA Diamond as a forward-looking differentiator.

MMMU-Pro / Video-MMMU:

- Gemini 3 Pro: MMMU-Pro 81.0%, Video-MMMU 87.6% — current leader.
- GPT-5.4 and Claude Opus 4.7 close behind on still-image multimodal; video remains a Google strength.

MMLU-Pro: Effectively saturated (>90% on top models); not reported separately.

5.5 Frontier-Hard Benchmarks

Humanity's Last Exam (HLE) — 2,500 questions; SEAL leaderboard updated April 19, 2026; 262+ models evaluated:

RANK	MODEL	ACCURACY (NO TOOLS)	ACCURACY (W/ TOOLS)
1	Claude Mythos Preview (restricted)	64.7%	—
2	Gemini 3 Deep Think (upgraded Feb 12, 2026)	48.4%	>50%
3	Gemini 3.1 Pro Preview	44.7%	—
4	GPT-5.4 xhigh	41.6%	—
5	GPT-5.3-Codex xhigh	39.9%	—
6	Claude Opus 4.6 Thinking Max	34.4%	40%
7	GPT-5 Pro	31.6%	—
—	Kimi K2.6 (with tools)	—	54.0%
—	Muse Spark Contemplating Mode	—	58%

Sources: Artificial Analysis HLE, llm-stats HLE, Price Per Token HLE.

The apparent tool-vs-no-tool delta (Kimi K2.6 at 54.0% with tools vs Gemini 3 Deep Think at 48.4% without) is genuinely informative: Kimi's 54.0% reflects the benefit of scaffolding and web search; Deep Think's 48.4% is a pure-model score and is the more demanding measurement.

ARC-AGI-1: Approaching saturation. Gemini 3 Deep Think 96%+ at $7.17/task; Opus 4.6 93.0% at $1.88/task; GPT-5.2 Pro 90.5% at $11.64/task — a 390× efficiency improvement over o3 ($4,500/task) in one year.

ARC-AGI-2: Gemini 3 Deep Think 84.6% (verified by ARC Prize Foundation) at $13.62/task. Gemini 3.1 Pro 77.1% (vendor). Claude Sonnet 4.6 58.3%, a 4.3× jump over Sonnet 4.5.

ARC-AGI-3 (NEW — launched March 27, 2026): ARC Prize Foundation released ARC-AGI-3, an interactive-reasoning benchmark of hundreds of turn-based environments. Frontier AI scores below 1% on ARC-AGI-3; humans score 100%. ARC-AGI-3 is the most important new benchmark released since February 2026. The ARC Prize 2026 competition has a >$2M prize pool spread across ARC-AGI-2 (Grand Prize guaranteed) and ARC-AGI-3 tracks.

5.6 Mathematical Reasoning

AIME 2025 — saturated. GPT-5.4 scores 100% with code execution; Gemini 3 Pro 100%; Deep Think 98–99%; Claude Opus 4.6 effectively solved. AIME 2025 is no longer a meaningful differentiator above 95%.

AIME 2026 (I and II, February 2026): Too new for a stable leaderboard snapshot; Gemini 3 Deep Think solved AIME 2026 #15 after ~30 minutes of thinking per published community traces.

USAMO 2026: GPT-5.4 95.2%; Deep Think gold-medal-level.

FrontierMath: GPT-5.4 47.6% (math leader per Vellum); Deep Think competitive on research-grade problems.

MathArena Apex: Gemini 3 Pro 23.4% — new state of the art vs mid-single-digits for prior frontier models.

IMO / IOI 2025: Gold-medal-level results from Gemini 3 Deep Think, GPT-5.3-Codex (derived from code reasoning), and DeepSeek V3.2.

06

Reading Frontier-Hard Benchmarks Without Fooling Yourself

The three benchmarks most reliably signal forward-looking progress; each must be read with explicit caveats.

Humanity's Last Exam: The state of the art among generally-available models is Gemini 3 Deep Think at 48.4% without tools — a jump of ~11 percentage points in ten weeks. Kimi K2.6's 54.0% with tools is not directly comparable. HLE has a private 500-question held-out set to guard against overfitting. Cost per task is the critical hidden variable: Gemini 3 Deep Think reportedly costs $13.62 per ARC-AGI-2 task and materially more on HLE given the deeper reasoning involved. Muse Spark's 58% HLE in Contemplating Mode is vendor-reported and should be re-verified when Artificial Analysis evaluations come online.

ARC-AGI-2: Gemini 3 Deep Think's 84.6% is the first verified-by-ARC-Prize score above 80% and was a major driver of the February→April progress. But ARC Prize's own "refinement" leaderboard shows that harness engineering (Poetiq's refinement on Gemini 3 Pro) moves baseline 31% → 54% at $31/task, suggesting that 10–20 points of measured ARC-AGI-2 progress across labs in Q1 2026 has been scaffolding-driven, not model-driven.

ARC-AGI-3: The new frontier benchmark. Frontier AI <1%, humans 100%. This benchmark will define "frontier" for the rest of 2026. No generally-available model has achieved >1% on the official (non-refinement) leaderboard as of April 23, 2026.

SWE-bench Pro: The coding benchmark least vulnerable to contamination. The public-private delta still averages 5–8 points. Scale AI's SEAL uncapped-cost, 250-turn standardized harness is the controlled comparison; agent-system scores with custom scaffolding are inflated by 4–10 points and are not cross-comparable. Cost per SWE-bench-Pro-class resolved task ranges from ~$0.23 (DeepSeek V3.2) to ~$1.06 (GPT-5.3-Codex with full Codex harness). These are directional only.

07

Methodological Cautions

Preserving Aiveris' six February cautions, updated with April 2026 evidence:

1. **Vendor-reported vs independent.** OpenAI's GPT-5.5 launch today is vendor-reported only; treat its "82.7% Terminal-Bench 2.0 state-of-the-art" as tentative until independent validation (tbench.ai, Scale SEAL, Artificial Analysis) arrives. Similar caveat for Anthropic's Opus 4.7 Terminal-Bench 2.0 69.4% claim — not yet on the tbench.ai public board.
2. **Scaffold inflation.** Morph's internal Opus 4.6 score on SWE-bench Pro ranges from 50.2–55.4% depending on scaffold — a 5-point swing from context management and retrieval quality alone. Terminal-Bench 2.0's top two scaffolds (ForgeCode + Opus 4.6 / GPT-5.4) both land at 81.8%, but the underlying model scores are 69–75%.
3. **Contamination.** OpenAI's internal audit disclosed that GPT-5.2, Opus 4.5, and Gemini 3 Flash can reproduce verbatim gold patches on SWE-bench Verified — proof that Verified is partly memorized. 59.4% of the hardest unsolved Verified tasks have flawed test cases. Lean on SWE-bench Pro's public-private delta and on SWE-rebench's post-release sourcing when precision matters.
4. **Saturation.** GPQA Diamond, AIME 2025, MMLU-Pro, and ARC-AGI-1 are all effectively saturated. The meaningful signal in Q2 2026 comes from HLE (no-tools), ARC-AGI-2, ARC-AGI-3, SWE-bench Pro, MathArena Apex, and long-horizon agentic evaluations (APEX-Agents, Vending-Bench 2, GDPval-AA).
5. **Composite indices hide the component.** Artificial Analysis Intelligence Index 57 is shared by Claude Opus 4.7, Gemini 3.1 Pro Preview, and GPT-5.4 xhigh — but Anthropic leads GDPval-AA Elo by ~439 points over Gemini 3.1 Pro at the same composite score. Always read component-level.
6. **Cost drift without price change.** Claude Opus 4.7's new tokenizer can increase effective cost by up to 35% at unchanged headline pricing. GPT-5.4's 272K→1M context tier doubles input cost. These are the types of changes that only appear on next month's invoice.

A new seventh caution emerged since February: **vendor-gated risk classifications.** GPT-5.3-Codex, Opus 4.7, and GPT-5.5 are being deployed with cybersecurity safeguards that can block legitimate security-research prompts. If your use case includes penetration testing or vulnerability research, verify eligibility for Anthropic's Cyber Verification Program or OpenAI's trusted-access program before committing to a model.

08

Synthesis and Deployment Recommendations

Category leadership summary (as of April 23, 2026)

- **Complex agentic coding:** Claude Opus 4.7 — first choice, priced at $5/$25 with 1M context. Fallback: GPT-5.3-Codex or GPT-5.5 (latter pending API).
- **Computer-use and desktop automation:** GPT-5.4 or GPT-5.5 — the only models past human-expert baseline on OSWorld.
- **Graduate-level science reasoning:** Gemini 3.1 Pro Preview ($2/$12) — best price-to-GPQA ratio; Gemini 3 Deep Think for research-grade problems where tool-free HLE matters.
- **Multimodal (image/video/audio):** Gemini 3.1 Pro, with Gemini 3 Flash for volume multimodal workloads.
- **General chat / human preference:** Claude Opus 4.7 Thinking leads Arena Text by a slim margin; Sonnet 4.6 is near-identical at 40% of the cost.
- **Long context (>500K tokens):** Opus 4.7 / Opus 4.6 / Sonnet 4.6 (1M, no premium); Gemini 3.1 Pro (1M); Grok 4.20 / 4.3 (2M, largest Western closed); Llama 4 Scout (10M, open).
- **Cost-efficient high-volume:** DeepSeek V3.2 ($0.28/$0.42), Kimi K2.6 ($0.95/$4), GLM-5.1 ($1.00/$3.20), GPT-5.4-mini ($0.40/$1.60), Haiku 4.5 ($1/$5).
- **Self-hosted / data-sovereign:** GLM-5.1 (MIT, SWE-bench Pro leader among open weights), Kimi K2.6 (modified MIT), Qwen 3.6 family (Apache 2.0), DeepSeek V3.2 (MIT).
- **Real-time / latency-first coding:** GPT-5.3-Codex-Spark on Cerebras (>1000 t/s); Mercury 2 (716 t/s per Artificial Analysis).
- **Image generation:** GPT Image 1.5 (AA LMArena leader at 1264 Elo); Gemini 3 Pro Image (1235 Elo); Flux 2; Seedream 4.5.

> **Routing recommendation for mixed enterprise workloads:** A three-tier router that sends 60–70% of traffic to a Sonnet 4.6 / Gemini 3 Flash / DeepSeek V3.2 tier, 20–30% to a flagship tier (Opus 4.7 / GPT-5.4 / Gemini 3.1 Pro) selected by task domain, and <5% to the reasoning-maximum tier (Opus 4.7 xhigh / GPT-5.4 Pro / Gemini 3 Deep Think) for genuinely high-stakes tasks, will typically reduce per-request cost by 50–75% against a flagship-default baseline while maintaining >95% of flagship-equivalent output quality. Back-test on your own traffic; effective cost-per-request can shift >20% with a single tokenizer change.

Procurement cautions specific to Q2 2026

- Anthropic has published no sunset date for Opus 4.6 as of Opus 4.7 launch; pooled rate limits allow mixed-version migration. Budget for a 0–35% cost increase per prompt under the 4.7 tokenizer.
- OpenAI's GPT-5.2 retires June 5, 2026; migration to 5.4/5.5 is mandatory.
- OpenAI's GPT-5.5 API access is "very soon" but not yet live (April 23, 2026); do not architect against 5.5-specific behavior until the API ships.
- Gemini 3.1 Pro remains in Preview — expect behavior changes.
- Meta Muse Spark has no published API rate card; it is currently in private API preview only.
- DeepSeek V4 has not shipped despite rumors; V3.2 remains the procurement target.

Category rule of thumb: If your procurement criterion is a single metric, the correct answer for April 2026 is usually "pick the model with the best component score in your domain, not the best composite." The top three composite-57 models diverge by up to 15 points on specific categories.

09

What This Analysis Cannot Tell You

No benchmark suite captures: (a) how a model behaves on your codebase, with your tool harness, your latency budget, and your non-English content mix; (b) reliability under adversarial and jailbreak pressure in your threat model; (c) the operational quality of a vendor's rate limits, region availability, incident response, and long-term SLA posture; (d) data-residency and regulatory posture for your jurisdiction (particularly relevant when evaluating Chinese open-weight options for regulated industries); (e) the human cost of model switching — integration, prompt library migration, eval harness rebuild, and team retraining; (f) the second-order effects of vendor-imposed capability safeguards on legitimate security-research, red-team, and compliance use cases; (g) future price, policy, or availability changes — OpenAI's API-tier pricing has shifted inside 72 hours of launch in past cycles; (h) whether a model announced today survives in production in six months.

Pilot, back-test on real traffic, and instrument for cost-per-successful-outcome (not cost-per-token).

10

Sources and Data Provenance

Official vendor announcements and documentation

Anthropic — "Introducing Claude Opus 4.7": https://www.anthropic.com/news/claude-opus-4-7

Anthropic — Claude Opus page: https://www.anthropic.com/claude/opus

Anthropic Platform — Pricing: https://platform.claude.com/docs/en/about-claude/pricing

Anthropic Platform — Release Notes: https://platform.claude.com/docs/en/release-notes/overview

Anthropic Platform — Models Overview: https://platform.claude.com/docs/en/about-claude/models/overview

OpenAI — Introducing GPT-5.5: https://openai.com/index/introducing-gpt-5-5/

OpenAI — Introducing GPT-5.4: https://openai.com/index/introducing-gpt-5-4/

OpenAI — Introducing GPT-5.3-Codex: https://openai.com/index/introducing-gpt-5-3-codex/

OpenAI — GPT-5.3-Codex-Spark: https://openai.com/index/introducing-gpt-5-3-codex-spark/

OpenAI — GPT-5.3-Codex System Card: https://cdn.openai.com/pdf/23eca107-a9b1-4d2c-b156-7deb4fbc697c/GPT-5-3-Codex-System-Card-02.pdf

OpenAI Developers — GPT-5.4 Model: https://developers.openai.com/api/docs/models/gpt-5.4

OpenAI — Model Release Notes: https://help.openai.com/en/articles/9624314-model-release-notes

Google — Gemini 3 launch blog: https://blog.google/products-and-platforms/products/gemini/gemini-3/

Google — Gemini 3.1 Pro blog: https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/

Google — Gemini 3 Deep Think blog: https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-deep-think/

Google DeepMind — Gemini 3.1 Pro model card: https://deepmind.google/models/model-cards/gemini-3-1-pro/

Google AI — Gemini API pricing: https://ai.google.dev/gemini-api/docs/pricing

Google AI — Gemini 3 Developer Guide: https://ai.google.dev/gemini-api/docs/gemini-3

Google Cloud — Vertex AI Gemini 3.1 Pro: https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/3-1-pro

Google Cloud — Agent Platform Pricing: https://cloud.google.com/vertex-ai/generative-ai/pricing

Meta AI — Introducing Muse Spark: https://ai.meta.com/blog/introducing-muse-spark-msl/

xAI — Grok Release Notes: https://grok.com/release-notes

Moonshot AI — Kimi K2.6 announcement: https://www.kimi.com/blog/kimi-k2-6

Z.ai — GLM-5.1 blog: https://z.ai/blog/glm-5.1

Z.ai — GLM-5.1 developer docs: https://docs.z.ai/guides/llm/glm-5.1

Hugging Face — zai-org/GLM-5: https://huggingface.co/zai-org/GLM-5

Hugging Face — zai-org/GLM-5.1: https://huggingface.co/zai-org/GLM-5.1

DeepSeek — V3.2 API docs: https://api-docs.deepseek.com/news/news251201

DeepSeek — V3.2-Exp: https://api-docs.deepseek.com/news/news250929

DeepSeek — Pricing: https://api-docs.deepseek.com/quick_start/pricing/

AWS — Claude Opus 4.7 in Bedrock: https://aws.amazon.com/blogs/aws/introducing-anthropics-claude-opus-4-7-model-in-amazon-bedrock/

GitHub — Claude Opus 4.7 in Copilot: https://github.blog/changelog/2026-04-16-claude-opus-4-7-is-generally-available/

Microsoft — GPT-5.4 in Foundry: https://techcommunity.microsoft.com/blog/azure-ai-foundry-blog/introducing-gpt-5-4-in-microsoft-foundry/4499785

Benchmark leaderboards

Arena (LMArena) — text: https://arena.ai/leaderboard/text/

Arena — leaderboard changelog: https://arena.ai/blog/leaderboard-changelog/

SWE-bench Leaderboards: https://www.swebench.com/

Scale AI SEAL — SWE-bench Pro (Public): https://labs.scale.com/leaderboard/swe_bench_pro_public

Scale AI SEAL — SWE-bench Pro (Private): https://labs.scale.com/leaderboard/swe_bench_pro_private

Scale AI SEAL — Humanity's Last Exam: https://labs.scale.com/leaderboard/humanitys_last_exam

SWE-rebench (Nebius): https://swe-rebench.com

SWE-bench Live: https://swe-bench-live.github.io/

Terminal-Bench 2.0: https://www.tbench.ai/leaderboard/terminal-bench/2.0

tbench.ai overview: https://www.tbench.ai/leaderboard

Artificial Analysis — Leaderboard: https://artificialanalysis.ai/leaderboards/models

Artificial Analysis — HLE: https://artificialanalysis.ai/evaluations/humanitys-last-exam

Artificial Analysis — Terminal-Bench Hard: https://artificialanalysis.ai/evaluations/terminalbench-hard

Artificial Analysis — GPT-5.4 model page: https://artificialanalysis.ai/models/gpt-5-4

Artificial Analysis — Gemini 3.1 Pro Preview: https://artificialanalysis.ai/models/gemini-3-1-pro-preview

Artificial Analysis — DeepSeek V3.2: https://artificialanalysis.ai/models/deepseek-v3-2

Artificial Analysis — Grok 4.20: https://artificialanalysis.ai/models/grok-4-20

ARC Prize: https://arcprize.org

ARC-AGI-3 announcement: https://arcprize.org/blog/arc-agi-3-launch

ARC-AGI-3: https://arcprize.org/arc-agi/3

ARC Prize 2025 results: https://arcprize.org/blog/arc-prize-2025-results-analysis

Humanity's Last Exam — CAIS: https://agi.safe.ai/

Epoch AI — Terminal-Bench 2.0: https://epoch.ai/benchmarks/terminal-bench

Epoch AI — HLE: https://epoch.ai/benchmarks/hle

Stanford HAI — 2026 AI Index: https://hai.stanford.edu/ai-index/2026-ai-index-report

Stanford HAI — Takeaways: https://hai.stanford.edu/news/inside-the-ai-index-12-takeaways-from-the-2026-report

Vellum — LLM Leaderboard: https://www.vellum.ai/llm-leaderboard

Price Per Token — GPQA: https://pricepertoken.com/leaderboards/benchmark/gpqa

Price Per Token — HLE: https://pricepertoken.com/leaderboards/benchmark/hle

Price Per Token — SWE-bench Lite: https://pricepertoken.com/leaderboards/benchmark/swe-bench-lite

BenchLM — SWE-bench Pro: https://benchlm.ai/benchmarks/swePro

BenchLM — Terminal-Bench 2.0: https://benchlm.ai/benchmarks/terminalBench2

BenchLM — SWE-rebench: https://benchlm.ai/benchmarks/sweRebench

llm-stats — SWE-Bench Verified: https://llm-stats.com/benchmarks/swe-bench-verified

llm-stats — SWE-Bench Pro: https://llm-stats.com/benchmarks/swe-bench-pro

llm-stats — Terminal-Bench 2.0: https://llm-stats.com/benchmarks/terminal-bench-2

llm-stats — HLE: https://llm-stats.com/benchmarks/humanity's-last-exam

OpenRouter — GPT-5.4: https://openrouter.ai/openai/gpt-5.4

OpenRouter — Gemini 3.1 Pro Preview: https://openrouter.ai/google/gemini-3.1-pro-preview

OpenRouter — Claude Haiku 4.5: https://openrouter.ai/anthropic/claude-haiku-4.5

OpenRouter — DeepSeek V3.2: https://openrouter.ai/deepseek/deepseek-v3.2

Journalism and independent analysis

Axios — Opus 4.7: https://www.axios.com/2026/04/16/anthropic-claude-opus-model-mythos

Axios — GPT-5.5 "Spud": https://www.axios.com/2026/04/23/openai-releases-spud-gpt-model

CNBC — Opus 4.7: https://www.cnbc.com/2026/04/16/anthropic-claude-opus-4-7-model-mythos.html

CNBC — Muse Spark: https://www.cnbc.com/2026/04/08/meta-debuts-first-major-ai-model-since-14-billion-deal-to-bring-in-alexandr-wang.html

CNBC — GPT-5.5: https://www.cnbc.com/2026/04/23/openai-announces-latest-artificial-intelligence-model.html

Fortune — GPT-5.3-Codex: https://fortune.com/2026/02/05/openai-gpt-5-3-codex-warns-unprecedented-cybersecurity-risks/

Fortune — Muse Spark: https://fortune.com/2026/04/08/meta-unveils-muse-spark-mark-zuckerberg-ai-push/

Fortune — GPT-5.5: https://fortune.com/2026/04/23/openai-releases-gpt-5-5/

TechCrunch — GPT-5.5: https://techcrunch.com/2026/04/23/openai-chatgpt-gpt-5-5-ai-model-superapp/

IEEE Spectrum — 2026 AI Index: https://spectrum.ieee.org/state-of-ai-index-2026

MIT Tech Review — 2026 AI Index charts: https://www.technologyreview.com/2026/04/13/1135675/want-to-understand-the-current-state-of-ai-check-out-these-charts/

VentureBeat — Muse Spark: https://venturebeat.com/technology/goodbye-llama-meta-launches-new-proprietary-ai-model-muse-spark-first-since

9to5Google — Gemini 3 Deep Think upgrade: https://9to5google.com/2026/02/12/gemini-3-deep-think-upgrade/

9to5Google — GPT-5.5: https://9to5google.com/2026/04/23/openai-releases-gpt-5-5/

Latent Space — Kimi K2.6: https://www.latent.space/p/ainews-moonshot-kimi-k26-the-worlds

Vellum — Gemini 3 benchmarks: https://www.vellum.ai/blog/google-gemini-3-benchmarks

WaveSpeed — Claude 4.6 family: https://wavespeed.ai/blog/posts/claude-opus-sonnet-4-6-whats-new/

WaveSpeed — Muse Spark vs Llama 4: https://wavespeed.ai/blog/posts/muse-spark-vs-llama-4-meta-strategy-2026/

WaveSpeed — GLM-5.1 comparison: https://wavespeed.ai/blog/posts/glm-5-1-vs-claude-gpt-gemini-deepseek-llm-comparison/

Caylent — Haiku 4.5: https://caylent.com/blog/claude-haiku-4-5-deep-dive-cost-capabilities-and-the-multi-agent-opportunity

Finout — Opus 4.7 tokenizer cost analysis: https://www.finout.io/blog/claude-opus-4.7-pricing-the-real-cost-story-behind-the-unchanged-price-tag

NxCode — GPT-5.4 guide: https://www.nxcode.io/resources/news/gpt-5-4-complete-guide-features-pricing-models-2026

Marco Patzelt — SWE-Bench leaderboard April 2026: https://www.marc0.dev/en/leaderboard

CodeAnt — SWE-bench Leaderboard: https://www.codeant.ai/blogs/swe-bench-scores

Morph — SWE-Bench Pro analysis: https://www.morphllm.com/swe-bench-pro

TokenMix — SWE-Bench April 2026: https://tokenmix.ai/blog/swe-bench-2026-claude-opus-4-7-wins

Smart Chunks — Artificial Analysis Intelligence Index April 2026: https://smartchunks.com/artificial-analysis-intelligence-index-april-2026-explained/

Ofox — LLM Leaderboard April 2026: https://ofox.ai/blog/llm-leaderboard-best-ai-models-ranked-2026/

Messenger Bot — Chatbot Arena 2026: https://messengerbot.app/chatbot-arena-explained-how-llm-leaderboards-actually-rank-ai-models-in-2026/

dsebastien — Kimi K2.6, Qwen, Gemma 4: https://www.dsebastien.net/kimi-k2-6-qwen-and-gemma-4-local-ai-is-catching-up/

BuildFastWithAI — Best AI Models April 2026: https://www.buildfastwithai.com/blogs/best-ai-models-april-2026

BuildFastWithAI — GLM-5.1 review: https://www.buildfastwithai.com/blogs/glm-5-1-open-source-review-2026

BuildFastWithAI — Qwen vs GLM vs Kimi 2026: https://www.buildfastwithai.com/blogs/qwen-3-6-plus-vs-glm-5-1-vs-kimi-2026

NYU Shanghai — GLM-5.1: https://rits.shanghai.nyu.edu/ai/glm-5-1-z-ais-open-weight-model-takes-1-on-swe-bench-pro/

IntuitionLabs — Claude pricing: https://intuitionlabs.ai/articles/claude-pricing-plans-api-costs

IntuitionLabs — Humanity's Last Exam: https://intuitionlabs.ai/articles/humanitys-last-exam-ai-benchmark

TechSifted — Grok 4.3 Beta: https://techsifted.com/posts/grok-4-3-review-april-2026/

Sebastian Raschka — Open-weight LLMs 2026 architectures: https://magazine.sebastianraschka.com/p/a-dream-of-spring-for-open-weight

This report was prepared by Aiveris, LLC, a Pennsylvania-registered AI products and services business, as a refresh of the February 16, 2026 comparative analysis. It reflects the publicly auditable state of frontier large language model capability, pricing, and ecosystem position as of April 23, 2026. All benchmark figures, model release dates, pricing tables, and strategic conclusions should be re-validated against primary vendor sources before commitment to procurement, routing, or deployment decisions. The ten-week interval since the prior edition included three flagship frontier releases (Opus 4.7, GPT-5.4, GPT-5.5, Muse Spark) and one new frontier benchmark (ARC-AGI-3); the next ten weeks are likely to include further iteration at comparable velocity.

AIVERIS

Frontier AI intelligence for decision-makers.

aiveris.ai · info@aiveris.ai

Prepared by Aiveris, LLC — a Pennsylvania-registered AI products and services company.