FREE BENCHMARK GUIDE · EVERY FRONTIER MODEL

The AI Model Benchmark Guide

28 frontier models compared on intelligence, agentic coding, cost per task and value, with straight recommendations by workload.

	RANK	INTELLIGENCE	CODING	PER TASK
Claude Fable 5.2 ANTHROPIC	#1	64.6	58.5	$26.54
GPT-6.2 Astra OPENAI	#2	62.1	53.5	$4.00
Gemini 4 Neon GOOGLE	#3	61.4	58.6	$6.44

In this guide

03 **The frontier at a glance**
Who leads, who surprises, and the numbers that matter

04 **Intelligence, ranked**
Artificial Analysis Intelligence Index v5.0.1, all 28 scored models

05 **Agentic coding**
The Coding Agent Index, each model in its own harness

06 **What a task actually costs**
Cost per Intelligence Index task, cheapest first

07 **Intelligence vs cost**
The value map, and the quadrant to buy from

08 **The AITR Value For Money Index**
Our own ranking: intelligence points per dollar

09 **The evals that compare**
Terminal-Bench 4.0, FrontierCode 2.0, HLE and GDPval, run by Artificial Analysis

10 **The Claude line-up**
Fable 5.2, Opus 5.5, Sonnet 5.5, Haiku 5.5, profiled

11 **The challengers**
The strongest non-Anthropic models and the best open weights

12 **Which model should you use?**
Straight recommendations by workload

13 **Pricing reference**
Every published rate, one table

14 **Sources & methodology**
Where every number comes from

How to read this guide. Bars are coloured by *maker*, not rank: the same lab keeps the same colour on every chart. Claude models carry a subtle accent so you can track them through the field. Nothing here is estimated: unpublished values are shown as “n/a”, and page 14 states the source for every data class.

The frontier at a glance

As of 22 November 2026, Claude Fable 5.2 leads the Artificial Analysis Intelligence Index v5.0.1 on 64.6, 2.5 points ahead of GPT-6.2 Astra. Gemini 4 Neon leads agentic coding in Antigravity CLI, and Muse Ember delivers the most intelligence per dollar.

MOST INTELLIGENT

64.6

Claude Fable 5.2, Intelligence Index v5.0.1

BEST CODING AGENT

58.6

Gemini 4 Neon in Antigravity CLI, Coding Agent Index

CHEAPEST PER TASK

$0.14

Muse Ember: lowest cost AA has measured

BEST VALUE

257

Muse Ember, intelligence points per dollar

TOP OPEN WEIGHTS

46.5

Kimi K3.1 Pro, #11 overall

THE PREMIUM

6.6×

Claude Fable 5.2 costs $26.54/task vs GPT-6.2 Astra's $4.00

The one-sentence verdict. Claude Fable 5.2 is the most capable model measured, but GPT-6.2 Astra delivers 96% of its score at 15% of the cost per task, and Kimi K3.1 Pro is the strongest model you can run on your own hardware.

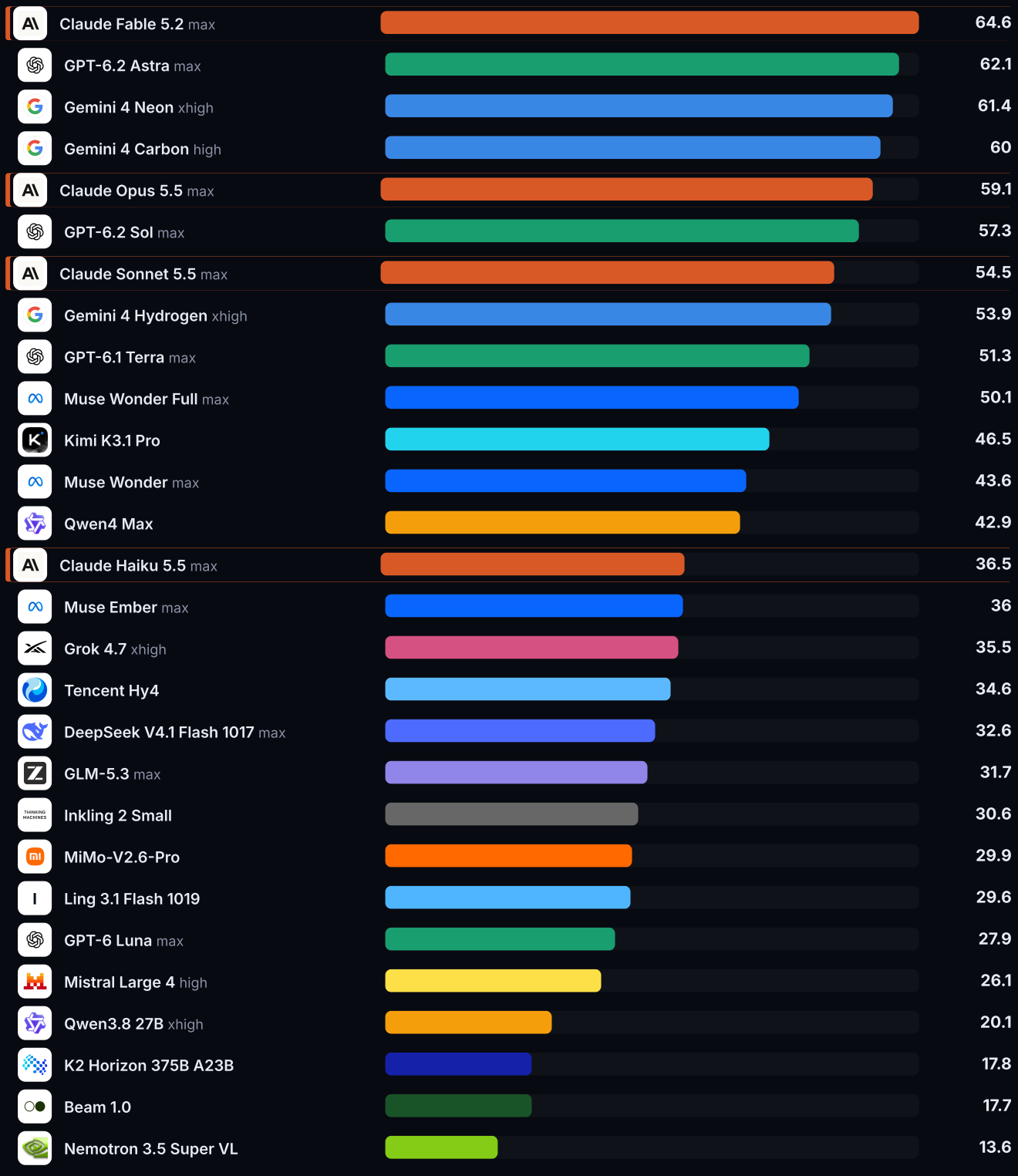
Intelligence, ranked

The Artificial Analysis Intelligence Index v5.0.1 blends many independently run evaluations (agentic work, coding, science, knowledge and reasoning) into one score. These are AA's own measurements, not vendor claims.

Anthropic · OpenAI · Google · Meta · Kimi · Alibaba · SpaceXAI · Tencent · DeepSeek · Z.ai · Thinking Machines · Xiaomi · InclusionAI · Mistral · Institute of Foundation Models · Reflection · NVIDIA

Artificial Analysis Intelligence Index v5.0.1

Higher is better · independent measurement, not vendor claims



Source: Artificial Analysis, 22 November 2026. Claude models carry the side accent.

Agentic coding

The Coding Agent Index runs each model inside a real coding agent (Claude in Claude Code, GPT in Codex, Gemini in Antigravity CLI and so on) on FrontierCode, SWE-Atlas-QnA and Terminal-Bench. It is the best predictor of real coding-agent performance.

Google · Anthropic · OpenAI · Kimi · Meta · SpaceXAI · Mistral · Z.ai · DeepSeek · Reflection



Source: Artificial Analysis, 22 November 2026. Gemini 4 Neon leads; Claude Fable 5.2 is 0.1 points back.

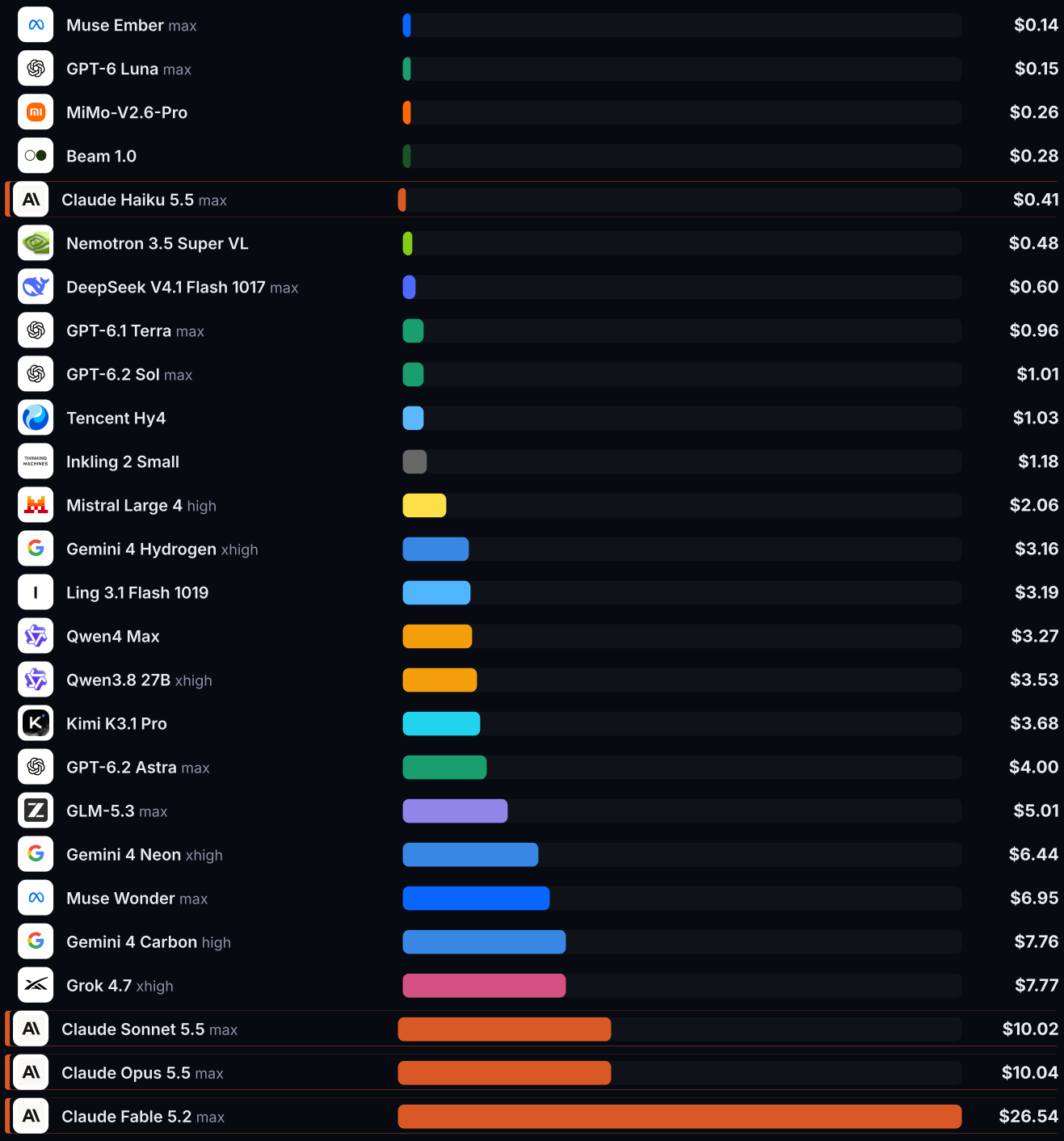
What a task actually costs

Per-token prices mislead, because reasoning models burn very different numbers of tokens. Artificial Analysis measures the weighted average cost to complete one Intelligence Index task, which is the number that hits your invoice.

Meta · OpenAI · Xiaomi · Reflection · Anthropic · NVIDIA · DeepSeek · Tencent · Thinking Machines · Mistral · Google · InclusionAI · Alibaba · Kimi · Z.ai · SpaceXAI

Cost per Intelligence Index task

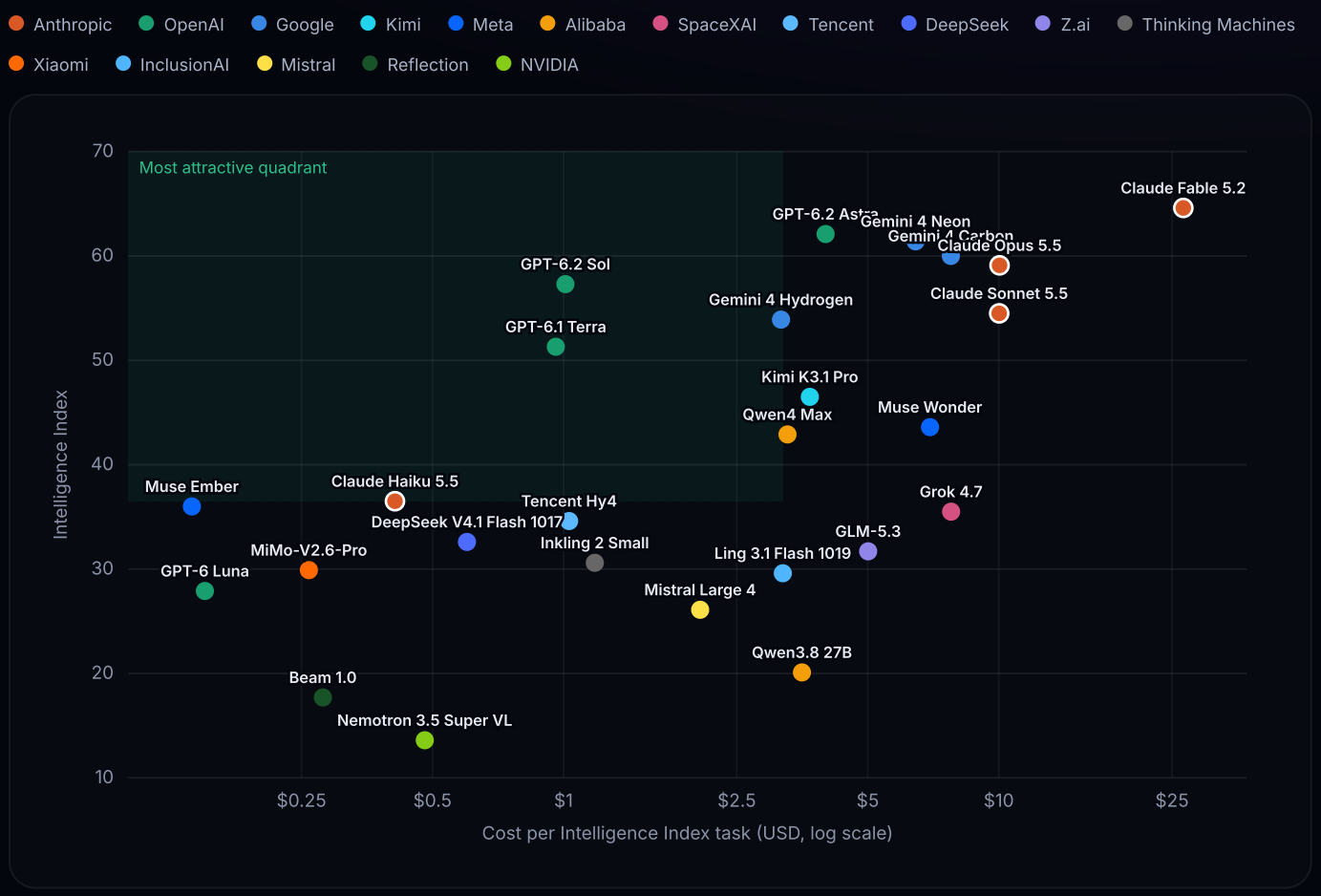
Lower is better · USD



Source: Artificial Analysis, 22 November 2026. The spread is 190×: $0.14 (Muse Ember) to $26.54 (Claude Fable 5.2).

Intelligence vs cost: the value map

Up and to the left is where you want to buy. The shaded quadrant holds above-median intelligence at below-median cost; Claude models are ringed in white.



Read it like this: GPT-6.2 Sol, Gemini 4 Hydrogen, GPT-6.1 Terra, Claude Haiku 5.5 sit in the attractive quadrant. Claude Fable 5.2 buys the top score at $26.54 per task: worth it only when that margin matters.

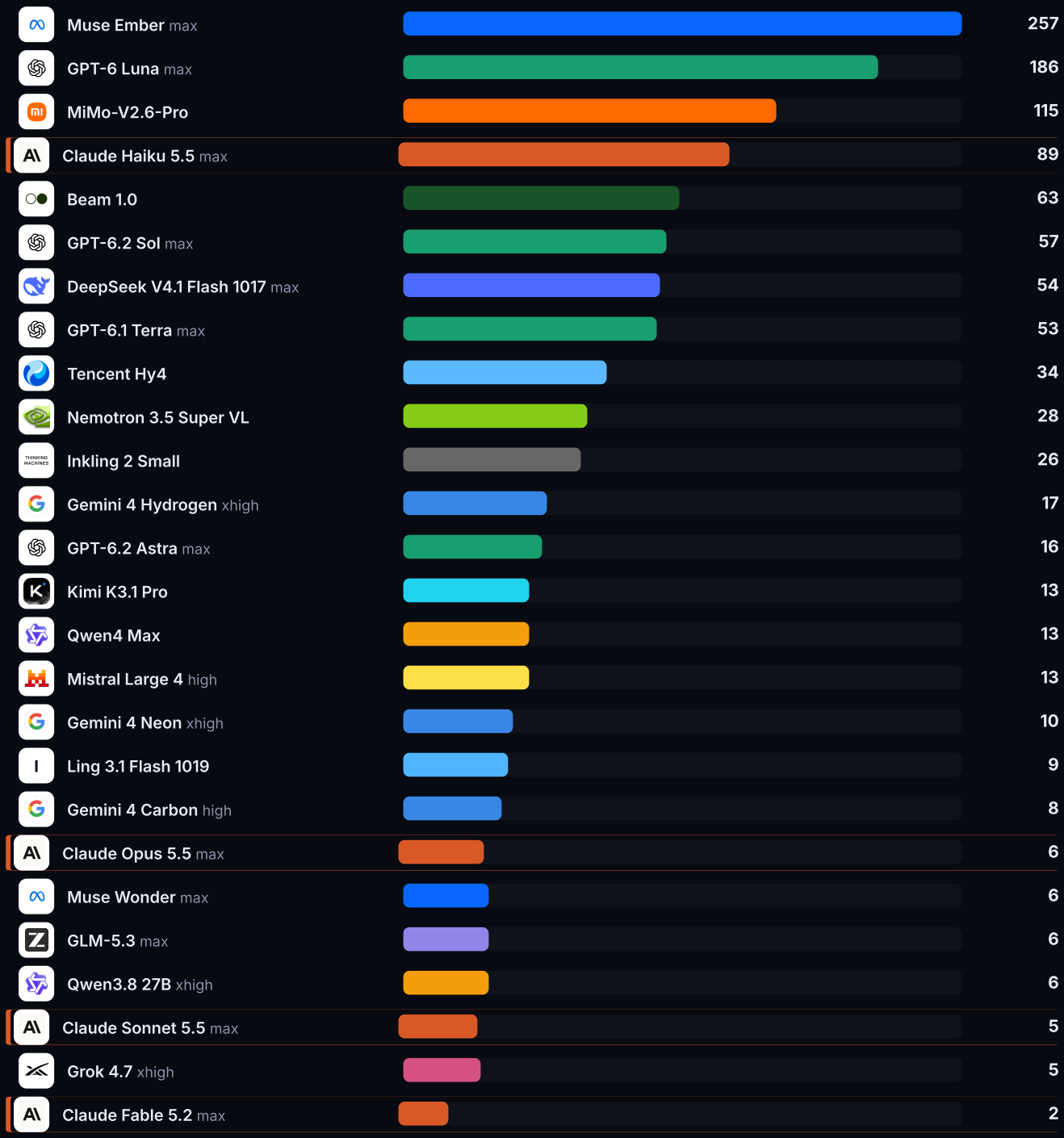
The AITR Value For Money Index

Our own derivation: Intelligence Index points per US dollar of cost per task. Simple declared arithmetic, and a ranking the headline charts hide.

Meta · OpenAI · Xiaomi · Anthropic · Reflection · DeepSeek · Tencent · NVIDIA · Thinking Machines · Google · Kimi · Alibaba · Mistral · InclusionAI · Z.ai · SpaceXAI

AITR Value For Money Index, intelligence points per dollar

Higher is better · bar length on a square-root scale for readability; figures are exact



Derived from Artificial Analysis intelligence and cost data (22 November 2026). Top three: Muse Ember, GPT-6 Luna, MiMo-V2.6-Pro.

The evals that compare across labs

Vendor launch tables use different harnesses, so they rarely compare. These four are run by Artificial Analysis itself, the same way for every model. Column leaders in bold.

MODEL	TERMINAL-BENCH 4.0	FRONTIERCODE 2.0	HUMANITY'S LAST EXAM	GDPVAL-AA (ELO)
Claude Fable 5.2	59.6%	**48.8%**	**74.3%**	1,831
GPT-6.2 Astra	64.1%	45.4%	n/a	n/a
Gemini 4 Neon	63.6%	44.4%	n/a	n/a
Gemini 4 Carbon	65.2%	40.6%	n/a	1,733
Claude Opus 5.5	59.6%	46.4%	61.4%	**1,866**
GPT-6.2 Sol	60.1%	44.4%	n/a	n/a
Claude Sonnet 5.5	63.6%	41.7%	55%	1,838
Gemini 4 Hydrogen	53.5%	39%	n/a	n/a
GPT-6.1 Terra	55.6%	42.3%	46.2%	1,476
Muse Wonder Full	57.1%	37.8%	53.6%	1,857
Kimi K3.1 Pro	**67.2%**	25.5%	52.5%	1,861
Muse Wonder	52%	27.9%	50.1%	1,770
Qwen4 Max	48.5%	26.5%	n/a	1,809
Claude Haiku 5.5	32.8%	34.9%	44.4%	1,629
Muse Ember	40.9%	27.2%	n/a	n/a
Grok 4.7	25.8%	28.7%	43.1%	1,717
Tencent Hy4	37.4%	31.4%	n/a	n/a
DeepSeek V4.1 Flash 1017	33.8%	28.5%	43.3%	1,746

Source: Artificial Analysis, 22 November 2026. Each model's own system-card results are on its page at aitoolsreview.co.uk/tool/<model>.

The Claude line-up

Claude Fable 5.2 max
ANTHROPIC

Intelligence Index	**64.6**
Coding Agent Index	**58.5 · Claude Code**
Cost per task	**$26.54**
API price (in/out per 1M)	**$8 / $40**
Context window	**1M**

#1 of 28 on the Intelligence Index v5.0.1 (64.6), Coding Agent Index 58.5, $26.54 per task. Fable 5.2 is the generally available configuration of a new Anthropic model released alongside Claude Mythos 5.2, which has more permissive cyber and biology safeguards and is restricted to vetted organisations.

Claude Opus 5.5 max
ANTHROPIC

Intelligence Index	**59.1**
Coding Agent Index	**58.2 · Claude Code**
Cost per task	**$10.04**
API price (in/out per 1M)	**$4 / $20**
Context window	**1M**

#5 of 28 on the Intelligence Index v5.0.1 (59.1), Coding Agent Index 58.2, $10.04 per task. Anthropic describes Claude Opus 5.5 as an upgrade to Opus 5 with gains in coding, agentic and computer use tasks, mathematical and scientific reasoning and long-horizon professional work.

Claude Sonnet 5.5 max
ANTHROPIC

Intelligence Index	**54.5**
Coding Agent Index	**56.3 · Claude Code**
Cost per task	**$10.02**
API price (in/out per 1M)	**$2 / $10**
Context window	**1M**

#7 of 28 on the Intelligence Index v5.0.1 (54.5), Coding Agent Index 56.3, $10.02 per task. Anthropic's system card finds Sonnet 5.5 broadly less capable than Opus 5.5 and not crossing any new RSP thresholds, with large gains over Sonnet 5 on most capability evaluations and a few healthcare results ahead of Opus 5.5.

Claude Haiku 5.5 max
ANTHROPIC

Intelligence Index	**36.5**
Coding Agent Index	**29.6 · Claude Code**
Cost per task	**$0.41**
API price (in/out per 1M)	**$0.1 / $0.5**
Context window	**1M**

#14 of 28 on the Intelligence Index v5.0.1 (36.5), Coding Agent Index 29.6, $0.41 per task. Anthropic's condensed system card finds Haiku 5.5 broadly less capable than Opus 5 and not crossing any new RSP thresholds, with large gains over Haiku 4.5 on every capability evaluation but usually below Sonnet 5.5.

The challengers

GPT-6.2 Astra max
OPENAI

Intelligence Index	**62.1**
Coding Agent Index	**53.5 · Codex**
Cost per task	**$4.00**
API price (in/out per 1M)	**$8 / $40**
Context window	**1M**

#2 of 28 on the Intelligence Index v5.0.1 (62.1), Coding Agent Index 53.5, $4.00 per task. OpenAI's system card rates GPT-6.2 Astra High for Biological and Chemical, Critical for Cybersecurity and High for AI Self-Improvement, the first externally released model to meet that threshold.

Gemini 4 Neon xhigh
GOOGLE

Intelligence Index	**61.4**
Coding Agent Index	**58.6 · Antigravity CLI**
Cost per task	**$6.44**
API price (in/out per 1M)	**$3 / $15**
Context window	**1M**

#3 of 28 on the Intelligence Index v5.0.1 (61.4), Coding Agent Index 58.6, $6.44 per task. Google DeepMind describes Gemini 4 Neon as a model adapted from Gemini 4 Carbon that balances depth and efficiency for complex coding, analytical reasoning and production enterprise workflows.

Gemini 4 Carbon high
GOOGLE

Intelligence Index	**60**
Coding Agent Index	**55.5 · Antigravity CLI**
Cost per task	**$7.76**
API price (in/out per 1M)	**$4 / $20**
Context window	**1M**

#4 of 28 on the Intelligence Index v5.0.1 (60), Coding Agent Index 55.5, $7.76 per task. Gemini 4 Carbon is Google DeepMind's top Gemini 4 model, based on Gemini 4 Argon and aimed at long-horizon agentic work, cybersecurity and scientific reasoning.

GPT-6.2 Sol max
OPENAI

Intelligence Index	**57.3**
Coding Agent Index	**49.8 · Codex**
Cost per task	**$1.01**
API price (in/out per 1M)	**$2 / $10**
Context window	**1M**

#6 of 28 on the Intelligence Index v5.0.1 (57.3), Coding Agent Index 49.8, $1.01 per task. GPT-6.2 Sol is OpenAI's cheaper everyday model, released with Astra at unchanged GPT-6.1 Sol prices.

Gemini 4 Hydrogen xhigh
GOOGLE

Intelligence Index	**53.9**
Coding Agent Index	**52.3 · Antigravity CLI**
Cost per task	**$3.16**
API price (in/out per 1M)	**$1.5 / $7.5**
Context window	**1M**

#8 of 28 on the Intelligence Index v5.0.1 (53.9), Coding Agent Index 52.3, $3.16 per task. Gemini 4 Hydrogen is Google DeepMind's fast, lower-cost Gemini 4 model and the successor to Gemini 3.8 Flash.

Kimi K3.1 Pro
KIMI

Intelligence Index	**46.5**
Coding Agent Index	**48.6 · Kimi Code CLI**
Cost per task	**$3.68**
API price (in/out per 1M)	**$2 / $10**
Context window	**1M**

#11 of 28 on the Intelligence Index v5.0.1 (46.5), Coding Agent Index 48.6, $3.68 per task. Open weights. Moonshot's model card reports Kimi K3.1 Pro at maximum reasoning effort (temperature 1.0, top-p 0.95) against Claude Fable 5.1, GPT-6 Astra, Claude Opus 5.5, DeepSeek-V4.1-Flash-1017 and GLM-5.3.

Which model should you use?

HARDEST REASONING, HIGHEST STAKES

Claude Fable 5.2

Top of the Intelligence Index v5.0.1 on 64.6. Budget $26.54 per task and use it where the last points of capability decide the outcome.

AGENTIC CODING AT SCALE

Gemini 4 Neon

Leads the Coding Agent Index on 58.6 in Antigravity CLI, at $6.44 per task.

NEAR-FRONTIER FOR LESS

GPT-6.2 Astra

96% of the leader's score at $4.00 per task, the cheapest model within 5 points of the top.

THE SENSIBLE DEFAULT

GPT-6.2 Sol

The highest-scoring model at or below the median cost per task ($3.19): 57.3 on the index for $1.01.

SELF-HOSTING / DATA CONTROL

Kimi K3.1 Pro

The strongest open-weight model measured (#11 overall). Run it on your own hardware or pay $3.68 per task.

HIGH-VOLUME PIPELINES

Muse Ember

The most intelligence per dollar measured: 257 points per dollar at $0.14 per task.

The routing pattern. Teams getting the most from the frontier don't pick one model. They route: a cheap tier for the bulk of calls, a mid model as the default, and the flagship reserved for the steps that genuinely need it.

Pricing reference

Every published rate in one place, as recorded by Artificial Analysis. Blended price weights input:output 3:1. "n/a" means no figure has been published: we never estimate.

MODEL	INPUT $/1M	OUTPUT $/1M	BLENDED	COST/TASK	CONTEXT
Claude Fable 5.2	$8	$40	$16.00	$26.54	1M
GPT-6.2 Astra	$8	$40	$16.00	$4.00	1M
Gemini 4 Neon	$3	$15	$6.00	$6.44	1M
Gemini 4 Carbon	$4	$20	$8.00	$7.76	1M
Claude Opus 5.5	$4	$20	$8.00	$10.04	1M
GPT-6.2 Sol	$2	$10	$4.00	$1.01	1M
Claude Sonnet 5.5	$2	$10	$4.00	$10.02	1M
Gemini 4 Hydrogen	$1.5	$7.5	$3.00	$3.16	1M
GPT-6.1 Terra	$1	$5	$2.00	$0.96	1M
Muse Wonder Full	n/a	n/a	n/a	n/a	1M
Kimi K3.1 Pro	$2	$10	$4.00	$3.68	1M
Muse Wonder	$4	$20	$8.00	$6.95	1M
Qwen4 Max	$1.5	$4.5	$2.25	$3.27	984k
Claude Haiku 5.5	$0.1	$0.5	$0.20	$0.41	1M
Muse Ember	$0.1	$0.5	$0.20	$0.14	1M
Grok 4.7	$2	$6	$3.00	$7.77	500k
Tencent Hy4	$0.834	$2.501	$1.25	$1.03	1M
DeepSeek V4.1 Flash 1017	$0.3	$1.2	$0.52	$0.60	1M
GLM-5.3	$1.4	$4.4	$2.15	$5.01	1M
Inkling 2 Small	$0.3	$1	$0.47	$1.18	1M
MiMo-V2.6-Pro	$0.435	$0.87	$0.54	$0.26	1M
Ling 3.1 Flash 1019	$0.3	$0.9	$0.45	$3.19	1M
GPT-6 Luna	$0.1	$0.5	$0.20	$0.15	1M
Mistral Large 4	$0.7	$2.1	$1.05	$2.06	1M
Qwen3.8 27B	$0.5	$3	$1.13	$3.53	256k
K2 Horizon 375B A23B	n/a	n/a	n/a	n/a	524k
Beam 1.0	$0.4	$1.2	$0.60	$0.28	1M
Nemotron 3.5 Super VL	$0.15	$0.5	$0.24	$0.48	1M

° observed maximum on OpenRouter serving endpoints rather than provider-claimed. Prices as of 22 November 2026.

Sources & methodology

Independent measurement. Intelligence Index v5.0.1, Coding Agent Index, cost per task and the four cross-lab evaluations are from Artificial Analysis (22 November 2026), which runs its evaluations itself against live provider APIs.

Vendor system cards. Model profiles draw on each vendor's own system card, model card or technical report (35 on file). Those figures are self-reported; the full tables are on each model's page.

Prices. Per-token prices are the list prices Artificial Analysis records for each model.

Observed serving data. Served context lengths are snapshotted from OpenRouter's public API (2026-11-22).

Our own derivation. The AITR Value For Money Index is Intelligence Index ÷ cost per task: declared arithmetic on the sourced inputs, with no editorial weighting.

Live version. This guide is regenerated from the same data as **aitoolsreview.co.uk/benchmarks**, which refreshes daily. Every model has its own page with its full system-card results at aitoolsreview.co.uk/tool/<model>.

© 2026 AIToolsReview · 22 November 2026