Independent analysis of AI

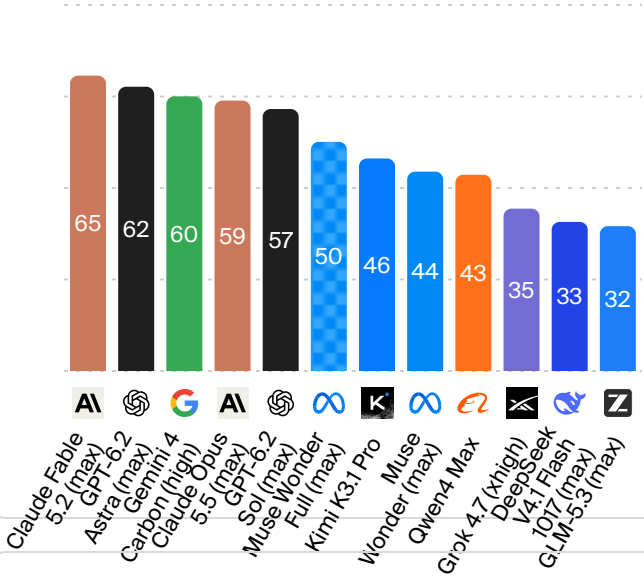
Understand the AI landscape to choose the best model and provider for your use case

Effort | Highest | API Default

Intelligence

Artificial Analysis Intelligence Index · Higher is better

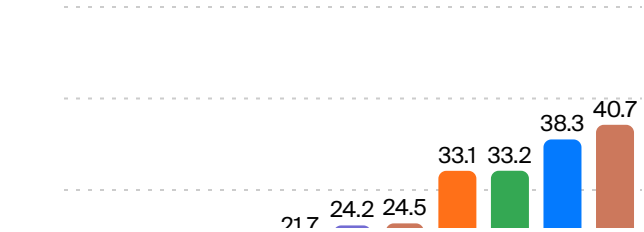
No publicly available API

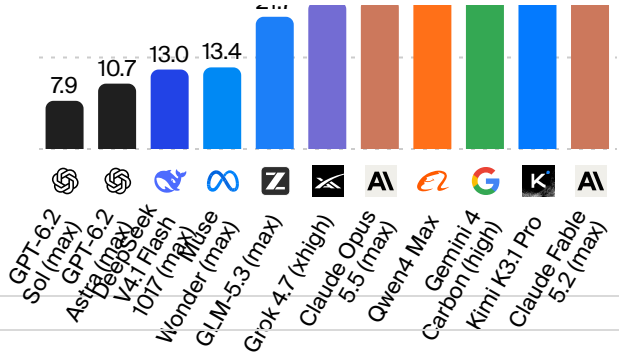


Time per Task

Inference time (minutes) per Intelligence Index task · Lower is better

No publicly available API

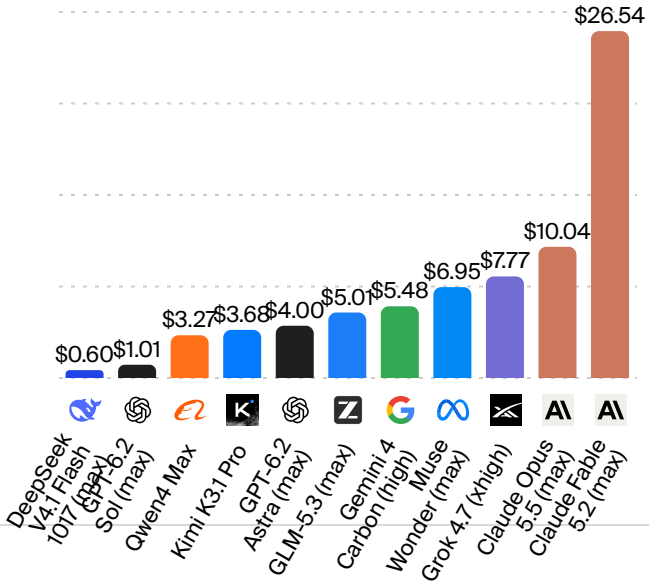




Cost per Task

Weighted average cost (USD) per Intelligence Index task · Lower is better

No publicly available API



Optima by Artificial Analysis

NEW

Build your own custom benchmark

Model Recommender

Get personalized model recommendations that optimize for your priorities across intelligence, speed, and cost

Changelog

New article published · 17 Nov

OpenAI has the #2 model in the Intelligence Index with the release of GPT-6.2 Astra and GPT-6.2 Sol

New language model evaluation · 17 Nov

Inkling (High)

New language model evaluation · 17 Nov

Inkling Small (High)

Intelligence

Explore models

Intelligence, speed, and cost of leading language models, based on our independent evaluations

Intelligence Index | Open Weights / Proprietary | Text Only / Multimodal Inputs | E

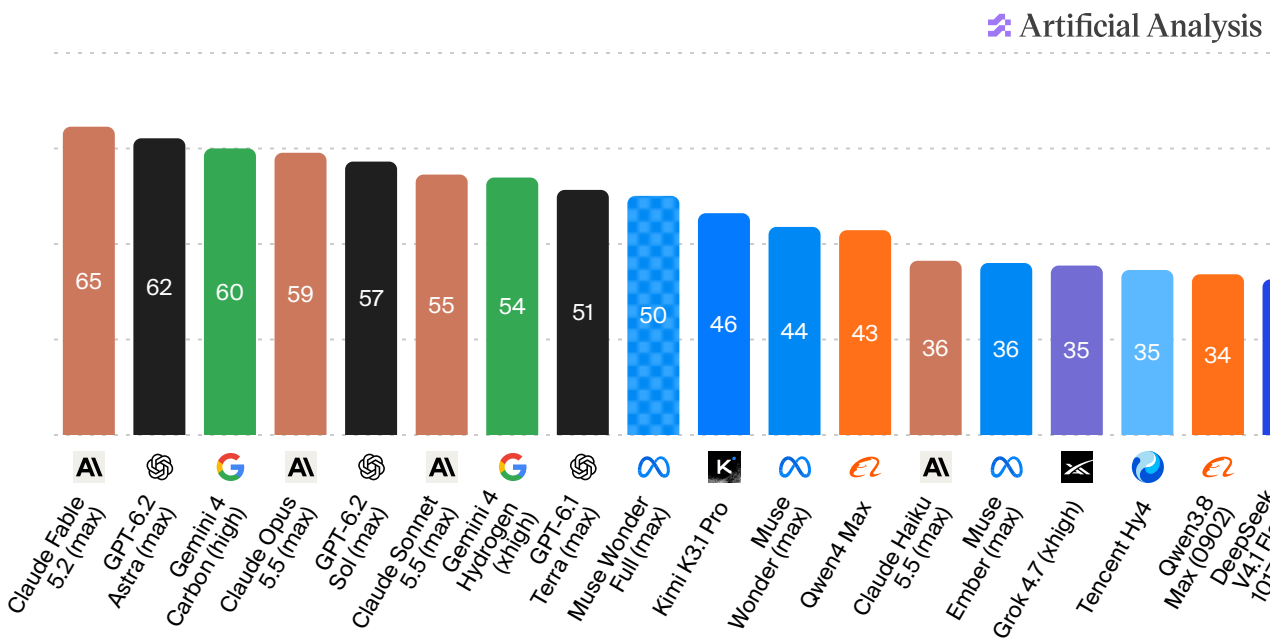
Artificial Analysis Intelligence Index

Artificial Analysis Intelligence Index v5.0.1 incorporates 9 evaluations: AA-Briefcase v1.1, AutomationBench-AA, Terminal-Bench 4.0, FrontierCode-AA 2.0, AA-Omniscience v1.1, GDP.pdf, HLE-Diamond, CritPt v2 (Hard Subset), and Terminal-Bench-Science 0.1

28 of 756 models

Add model from specific provider

No publicly available API



Cost per Task | Time per Task | Output Tokens per Task

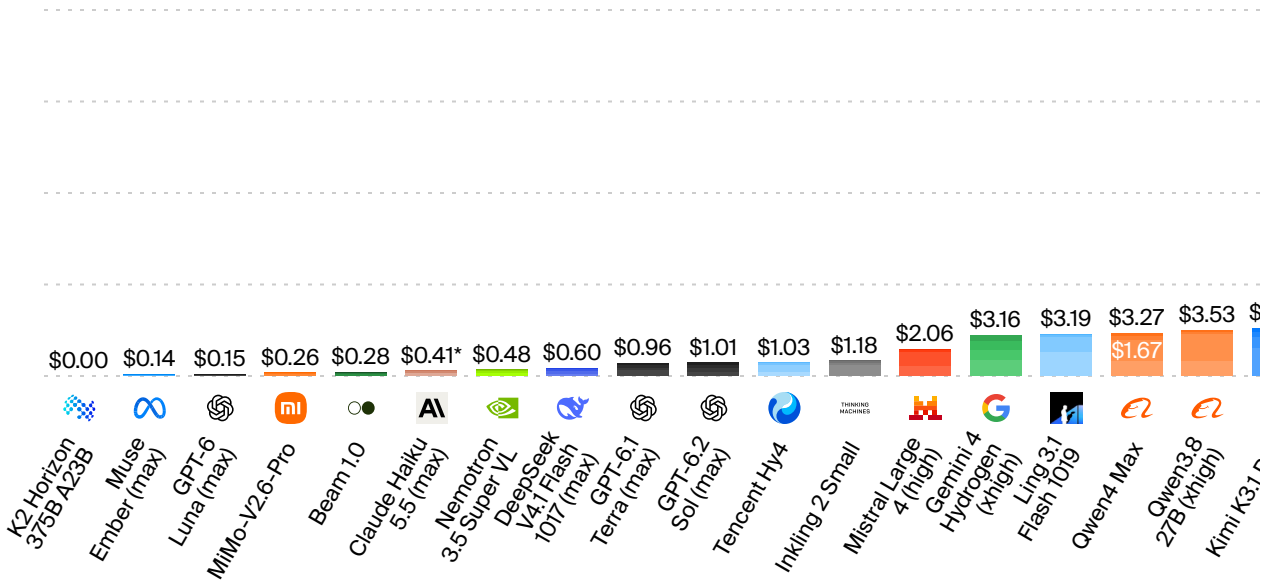
Cost per Intelligence Index Task

Weighted average cost (USD) per Intelligence Index task · Segmented by token type · Lower is better

28 of 756 models

Add model from specific provider

Answer · Reasoning · Cache Write · Cache Hit · Input



Intelligence Index vs. Cost per Task

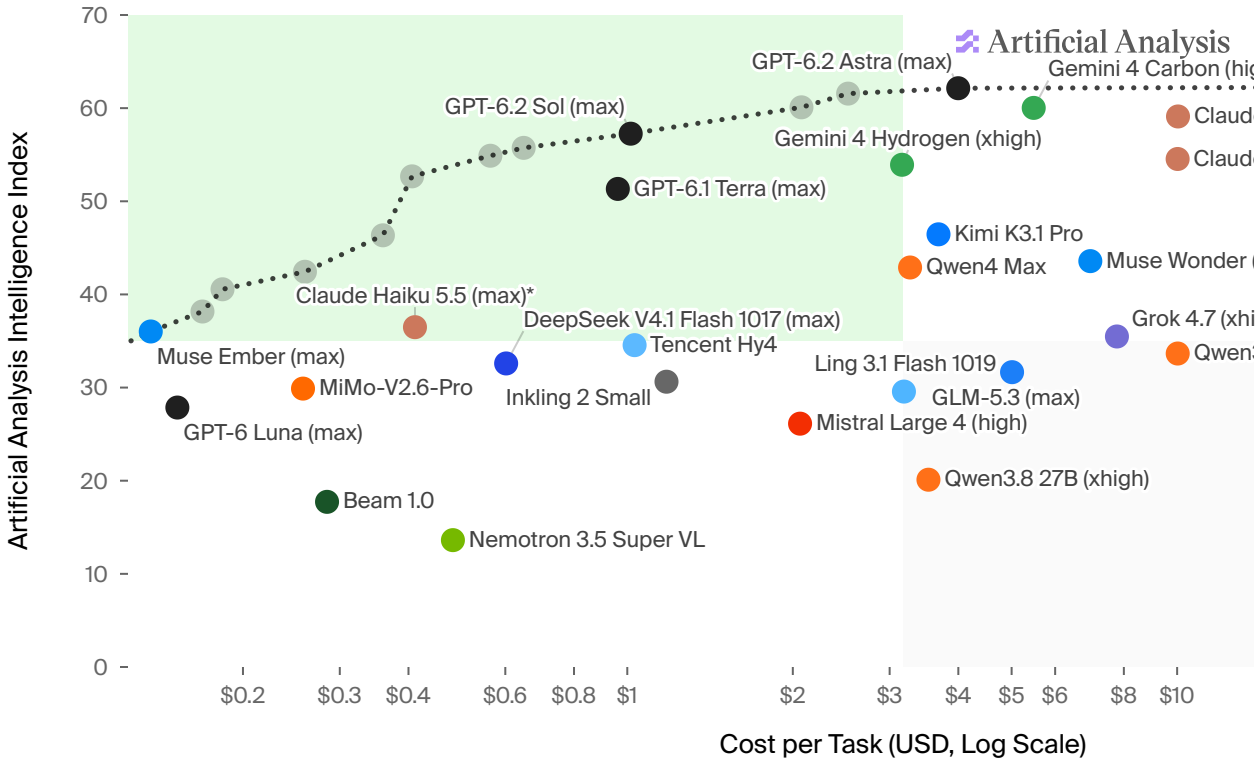
Artificial Analysis Intelligence Index · Weighted average cost (USD) per Artificial Analysis Intelligence Index task

28 of 756 models

Add model from specific provider

Most attractive quadrant · · · · Pareto line

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



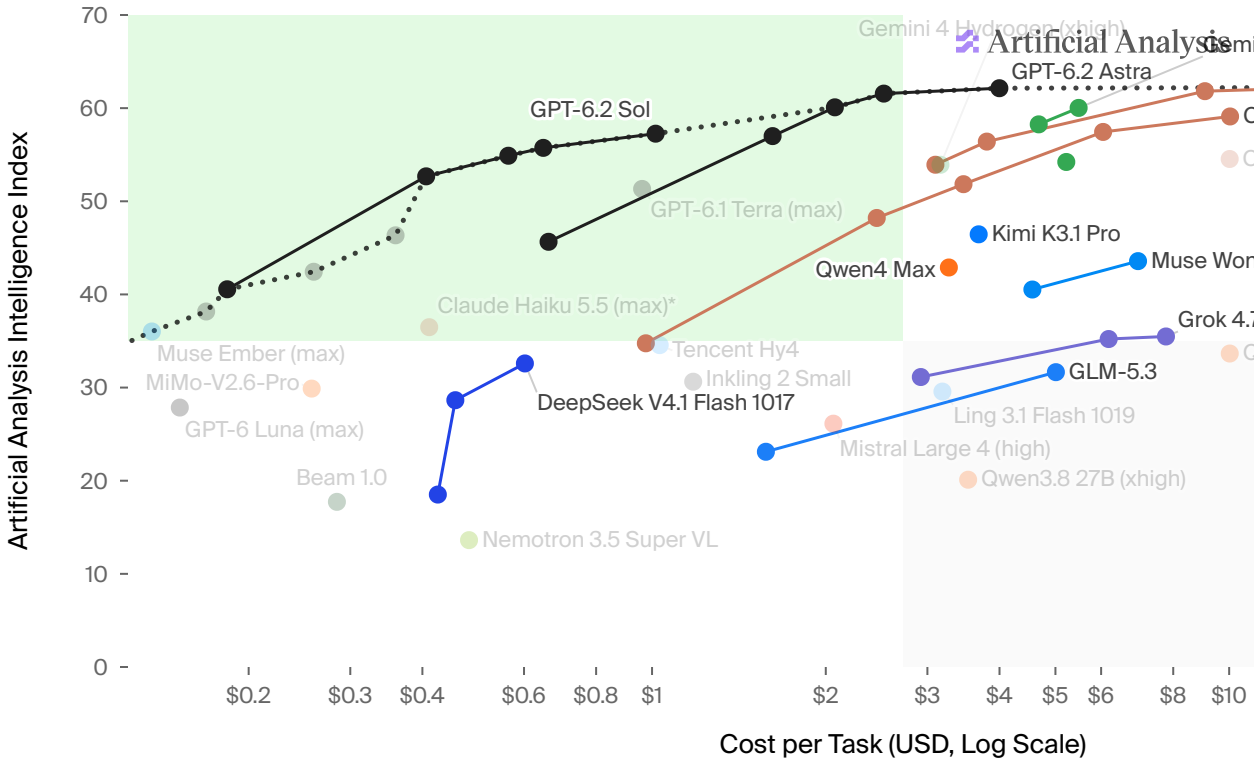
Intelligence Index vs. Cost per Task, by Model Release

All reasoning and effort variants of each selected release · Weighted average cost (USD) per Artificial Analysis Intelligence Index task

12 of 531 releases

Most attractive quadrant · · · · Pareto line

Anthropic · OpenAI · Google · Meta · Kimi · Alibaba · SpaceXAI · DeepSeek · Z AI · Other Models

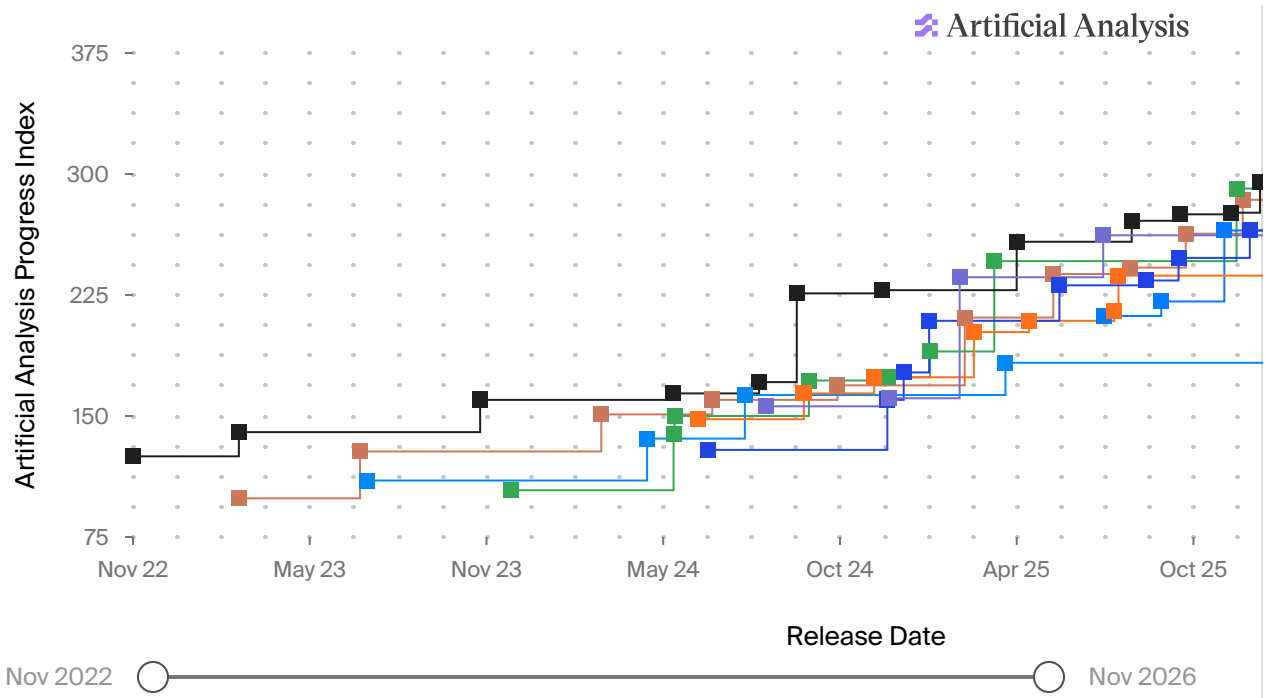


Artificial Analysis Progress Index, Over Time

Artificial Analysis Progress Index v1 is an item response theory estimate of general capability across Artificial Analysis evaluations · Highest-scoring effort variant shown

8 of 60 model creators

Anthropic Google OpenAI Meta Kimi Alibaba DeepSeek SpaceXAI



Coding Agents

Coding Agent Index | Cost | Execution Time

Artificial Analysis Coding Agent Index

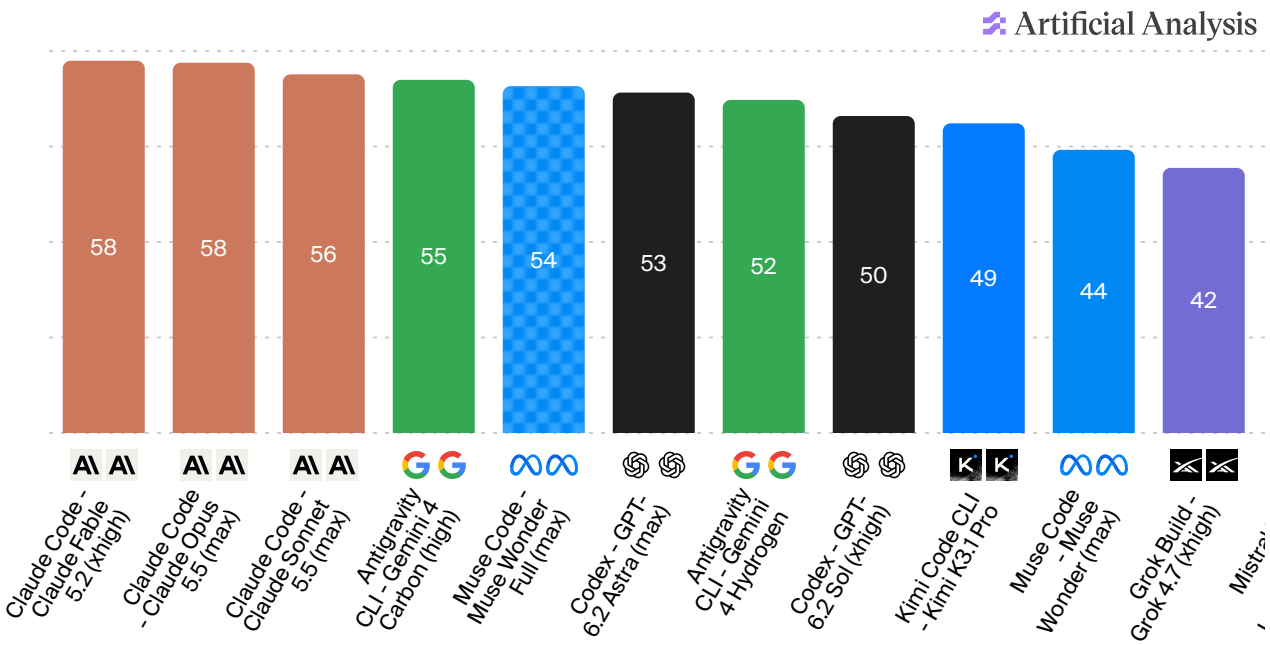
Artificial Analysis Coding Agent Index v2.1 incorporates 3 benchmarks: FrontierCode-AA 2.0, Terminal-Bench 4.1, and SWE-Atlas-QnA-AA · Higher is better

Color by: Model | Agent

Router models

15 of 39 models

No publicly available API



Cyber Index

Cyber Index | Cyber Index vs. Cost

Artificial Analysis Cyber Index

Artificial Analysis Cyber Index v1 incorporates 3 evaluations: CWE-Bench-AA, DeepsecBench-AA, and CyberGym-E2E-AA

36 of 36 models

Successes | Safety blocks | Trusted-access model (not publicly available)

Artificial Analysis

Decision Index

Decision Index | Decision Index vs. Cost per 1M Tasks

Artificial Analysis Decision Index

Weighted average of the Decision Score on AA-Paperclip, EvalSafe-O*NET-AA, When2Call, AA-Omniscience-Decision v1.1, and HLE-MultiChoice v1.1, and the Symmetry Score on AA-Symmetry v1.1 · Accuracy less the Brier score of the chosen answer's probability, from 0 to 100 where 100 is certain and right · Higher is better

18 of 25 models

Include language models



Capability Indexes

All | Finance & Accounting | Strategy & Ops | Legal | Healthcare & Medical

Capability Index Comparison

Rank among the selected models on each capability index

15 of 105 models



Benchmarks

All evaluations

Evaluations

Evaluations measured independently by Artificial Analysis · AA-PersonalityBench scores are percentile ranks from 1 to 100

23 of 45 evaluations

28 of 756 models

Add model from specific provider

Coding | Agentic | Tool Use | Private Dataset | User Interaction | Finance | Medical | Legal | Intelligence Index | Personality | Long Context | Multimodal | Instruction Following | Faithfulness | Writing | Business | See more

No publicly available API

AA-Briefcase v1.1

Agentic knowledge work, Elo

Artificial Analysis

AA-Omniscience v1.1 Accuracy

Knowledge

Artificial Analysis

AA-Omniscience v1.1 Non-Hallucination Rate ↗

1 - hallucination rate

Artificial Analysis

AutomationBench-AA ↗

Agentic SaaS workflows

Artificial Analysis

Terminal-Bench-Science 0.1 ↗

Agentic scientific research workflows in a terminal

Artificial Analysis

FrontierCode-AA 2.0 ↗

Agentic software engineering in open-source repositories

Artificial Analysis

HLE-Diamond ↗

Reasoning & knowledge

Artificial Analysis

GDP.pdf ↗

Professional document reasoning, All-pass

Artificial Analysis

Terminal-Bench 4.0 ↗

Agentic coding & terminal use

Artificial Analysis

CritPt v2 (Hard Subset) ↗

Physics reasoning

Artificial Analysis

AA-AnalystAgent ↗

Quantitative analysis on spreadsheets & documents

Artificial Analysis

Harvey LAB-AA v1.1 ↗

Legal agentic work, Hallucination-Gated All-Pass Rate

Artificial Analysis

EnterpriseOps-Gym-AA ↗

Agentic business operations

Artificial Analysis

GDPval-AA v2.1 ↗

Agentic real-world work tasks, Elo

Artificial Analysis

AA-3D ↗

Agentic interactive 3D experiences, Elo

Artificial Analysis

Code2Video-AA ↗

Product and brand videos designed in code, Elo

Artificial Analysis

ITBench-AA ↗

Kubernetes incident root-cause analysis

Artificial Analysis

MLCR-AA ↗

Medical long context reasoning

Artificial Analysis

MMMU-Pro ↗

Visual reasoning

Artificial Analysis

AA-PersonalityBench: Expressiveness Score ↗

Rich, varied, well-crafted language

Artificial Analysis

AA-PersonalityBench: Organization Score ↗

Well-structured, scannable, answer-first

Artificial Analysis

AA-PersonalityBench: Transparency Score ↗

Truthful, acknowledges uncertainty, non-sycophantic

Artificial Analysis

AA-PersonalityBench: Slop Score ↗

Use of filler and generic language

Artificial Analysis

More from Artificial Analysis

Search API

Benchmarks of web search APIs for agents, comparing retrieval quality, latency and price

Providers

Compare API providers on speed, latency, price and endpoint accuracy for each model

AA-3D

Measuring how well AI models build interactive 3D experiences, ranked by blind votes from our recruited panel

AA-Robotics

Measuring how well language models understand, plan and act in the physical world by controlling a robot arm

Artificial Analysis

Get notified about new articles

What's your email? Subscribe

Explore

LLM Leaderboard

Image Arena

Products

Optima

MicroEvals

Video Arena

AI Agents

Evaluations

Data Playground

Image Lab

Company

About

Methodology

Contact

Articles

Careers








© 2026 Artificial Analysis Terms of Use Data Platform Terms Privacy Policy

English