

Artificial Analysis Al Review

2024 Highlights

Artificial Analysis is a leading and independent AI benchmarking and insights provider. We support engineers and companies in understanding AI capabilities and in making critical decisions about their AI strategies.

Our data, insights and publications are grounded in our comprehensive benchmarking of AI technologies and use cases. This includes everything from hourly performance testing of language model APIs to millions of votes in our crowd-sourced evaluations.

Our public website, <u>artificialanalysis.ai</u>, is widely referenced throughout the AI industry. To discuss this report, our publications or our services, please contact us at <u>contact@artificialanalysis.ai</u>.



FRONTIER MODELS

2024 saw multiple labs catch up to OpenAI's GPT-4, and the emergence of the first models to push beyond GPT-4's level of intelligence

Frontier Language Model Intelligence, Over Time¹



Key Trends

Competing labs catch up to OpenAl's GPT-4: OpenAl started the language model race in Nov 2022 with the launch of GPT-3.5 in ChatGPT; competing labs spent the following 18 months catching up

Open models close in on the frontier labs: Open weights models, led by those from Meta, Mistral and Alibaba, have approached and surpassed GPT-4's level of intelligence

Sparks of intelligence beyond GPT-4: The final months of 2024 have seen the emergence of the first major intelligence leaps beyond GPT-4, led by OpenAI's o1. Topics including inference-time compute scaling, data quality and new reinforcement learning techniques have joined pretraining compute scaling as dominant levers for improving models

1. Artificial Analysis Quality Index: average across a range of language model intelligence and reasoning evaluation datasets. Currently includes MMLU, GPQA Diamond, MATH-500 & HumanEval. Estimated based on lab-claimed evaluation results for some non-current models. Release date is based on first public launch of the model . 2. Artificial Analysis Quality Index has been estimated based on lab-claimed evaluation results for certain models (non-current or API not yet available).



LANGUAGE MODEL COUNTRY OF ORIGIN

The US dominates the intelligence frontier; China appears to be in a clear second place and only a handful of other countries have demonstrated frontier-class training

The Language Model Frontier: Country of Origin

Artificial Analysis Quality Index, Leading Models (Late 2024)





OPEN-SOURCE LANGUAGE MODEL INTELLIGENCE

Driven by models from Meta, Mistral and Alibaba, the performance gap between open source and proprietary models has decreased significantly

Model Quality: Leading Proprietary and Open Weights Models

Based on proprietary and open-source models that resulted in an increase in Artificial Analysis Intelligence Index score





LANGUAGE MODEL INFERENCE PRICING

Language model inference pricing fell dramatically in 2024 for all levels of intelligence; GPT-40 mini approaches GPT-4 intelligence at a 100x cheaper price



1 Artificial Analysis Quality Index: average across a range of language model intelligence and reasoning evaluation datasets. Currently includes MMLU, GPQA Diamond, MATH-500 & HumanEval.



LANGUAGE MODEL SIZE

A key driver of the decline in inference pricing and increase in speed has been small models achieving levels of intelligence previously reserved for larger models **Quality by Model Size Over Time**

Highest Intelligence Index for each model parameter count; 3Q23 to 4Q24¹



Key takeaway

Over the last 12 months, small models have improved significantly. This **rate of improvement has significantly outpaced larger models**, resulting in a substantial narrowing of the performance gap to the frontier

What's Driving the Improvement?





Context windows have increased significantly to 128k as the new norm; long context reasoning enables models to work with more data at once

Context Length (Tokens), Median by Quarter

Median context length (thousand tokens)¹



The median context length of frontier models has increased 32x since 3Q23. Expansion was originally led by proprietary models but as of 3Q24 open source has caught up. Recent advancements have also increased the maximum context length of some models (Gemini, Nova) to 2M tokens

What's Driving the Improvement?

New Techniques	• Techniques for increasing context length have ranged from hardware aware distributed attention implementations to attention approximation and length extrapolation methods such as RoPE				
Developer Demand	 70% of developers in the Artificial Analysis Developer Survey indicated the context window was important to them when choosing a model Labs are responding to a clear industry demand signal 				
Implications					
Reduced Complexity	 Managing smaller context windows in production leads to complexity and tradeoffs Long contexts reduce your need to establish retrieval, summarization and truncation strategies 				
New Approaches & Applications	 Tools for agents quickly deplete the context window Many applications don't leverage only text. Larger context windows support multi-modal inputs including images, video, and audio 				
Inference Strategies	• There is a growing trend of more compute being utilized per task, whether in the form of generating additional reasoning tokens or agentic workflows. Longer contexts enable more flexibility				



LANGUAGE MODELS: KEY PLAYERS LANSCAPE

Players in the AI value chain differ in levels of vertical integration; Google stands out as the most vertically integrated player from TPU accelerators to Gemini





Demand for AI models is concentrated on releases from top AI labs; model reasoning quality and price are the primary decision drivers for choosing models

Model Demand by Provider

Which LLMs are you using or considering using?, N=270



Importance of Model Decision Criteria

How important are these criteria to you when choosing a model?, N=250

	Not important	Less Important	Important	Very important
Reasoning quality	0%	2.8%	32.5%	64.7%
Embedded knowledge	3.7%	21.5%	37.2%	37.6%
Context window	3.6%	20.6%	47.4%	28.3%
Speed / Throughput (Tokens	3.2%	21%	39.9%	35.9%
Latency (Time to First Token)	4.9%	27.6%	38.3%	29.2%
Price	2.4%	14.4%	33.2%	50%
Open- source	19.4%	31.2%	26.7%	22.7%
Function calling	8.5%	27.2%	38.6%	25.6%
JSON mode	13.4%	27.6%	34.6%	24.4%

Note: Results from the Artificial Analysis Developer Survey conducted from March to August 2024. Respondents represented a range of organization sizes and locations. Results should be considered indicative only and may be biased by Artificial Analysis's audience. Results may also be affected by survey timing and when new models were added to the survey (typically within days of their release).



LANGUAGE MODEL DEVELOPER INSIGHTS

Artificial Analysis Developer Survey

Companies are using LLMs in a wide range of technical approaches with no single approach dominant; most LLM users intend to use multimodal capabilities

Adoption of Technical Approaches for Using LLMs What technical uses do you intend to use LLMs for?, N=242 Information retrieval & large document(s) summarization (RAG, e.g. supply car manual and ask what an error light means) 177 resp. 68.3% Text-generation (e.g. email generation, generating text to respond to audio questions) 174 resp. 67.2% 138 resp. 53.3% Summarization (limited input, e.g. summarizing news articles) Chatbots (non-RAG, limited input) 124 resp. 47.9% Text structuring (e.g. turn this text into JSON with the following 119 resp. 45.9% structure) Classification (e.g. sentiment analysis) 117 resp. 45.2%

Demand for Multimodal Capabilities

What other AI capabilities do you use or intend to use?, N=252



Note: Results from the Artificial Analysis Developer Survey conducted from March to August 2024. Respondents represented a range of organization sizes and locations. Results should be considered indicative only and may be biased by Artificial Analysis's audience and impacted by survey timing.



LANGUAGE MODEL DEVELOPER INSIGHTS

Artificial Analysis Developer Survey

How are you hosting or considering hosting LLMs?, N=254

The vast majority users of AI models intend to use more than a single model in their applications; ~3/4 of users are accessing models via hosted serverless endpoints Number of Models Used Model Access & Deployment Approaches

How many different models do you use or plan to use?, N=265



Note: Results from the Artificial Analysis Developer Survey conducted from March to August 2024. Respondents represented a range of organization sizes and locations. Results should be considered indicative only and may be biased by Artificial Analysis's audience and impacted by survey timing.



IMAGE GENERATION QUALITY

Image generation quality progressed rapidly in 2024, with notable leaps in photorealism, prompt adherence, and text rendering

Image Generation Quality, Over Time

Prompt: The words 'Artificial Analysis' emblazoned on a next-generation spacecraft orbiting a breathtaking Earth view



1. Artificial Analysis Image Arena ELO, calculated as of 16 December 2024 based on 1.6 million votes from Artificial Analysis users

Artificial Analysis

IMAGE GENERATION LANSCAPE

Progress and competition for image models accelerated over the course of 2024; The top 5 models in the Artificial Analysis Image Arena all launched since Q3 2024

Image Arena ELO vs. Model Release Date

Release Arena Creator Model **ELO** Date 1.200 Recraft V3 Recraft Al 1.150 FLUX1.1 [pro] Stability.Al Ideogram v1 FLUX.1 [pro] R **Recraft Al** OCT 2024 Recraft v3 1161 ldeogram v2 1.100 Black Forest Labs Midjourney v6.1 🗨 🖕 Midjourney v6 Ideogram v2 Turbo Ideogram 1.050 Stable Diffusion 3 Large Recraft 20B Playground v3 (beta) Amazon DALLE 3 HD A **Black Forest labs** FLUX1.1 [pro] OCT 2024 1.000 1128 Playground v2.5 • Adobe Firefly 3 Playground AI 950 Amazon Titan G1 (Standard) Midjourney Amazon Titan G1 v2 (Standard) Stable Diffusion 1.6 Adobe 900 **Black Forest labs** FLUX.1 [pro] 1110 SDXL Lightning A AUG 2024 OpenAl Stable Diffusion XL 1.0 850 800 Midjourney Midjourney v6.1 JUL 2024 1092 750 DALLE 2 Stable Diffusion 2.1 700 ______3 Ideogram Ideogram v2 AUG 2024 1088 650 Stable Diffusion 1.5 600 04/22 07/22 10/22 01/23 04/23 07/23 01/24 04/24 07/24 10/24 01/25 10/23

Image Arena Leaderboard – Top 5 Models

About the arena

Arena ELO Score¹

Model Release Date

The Artificial Analysis Image Arena uses crowdsourced preferences to assess text to image model quality. As of December 2024, over 1.5 million preferences have been gathered. These preferences are used to compute Arena ELO scores, which reflect the relative quality of image models.

1. As at 4 October 2024. 2. FLUX 1.1 [pro] was tested on the Image Arena during 3Q24 under the pseudonym Blueberry



VIDEO GENERATION LANSCAPE

OpenAI previewed Sora in February 2024 to very little competition, but by the time it launched in December 2024, it joined a much more crowded field



Creator Model **ELO** Date 1.150 • Kling 1.5 Sora Hailuo AI • 1.100 Hunyuan Video K OpenAl Sora DEC 2024 1147 Runway Gen 3 Alpha 1.050 Kling 1.0 Haiper 2.0 • Mochi 1 Luma Dream Machine 1.000 Pika 1.5 80 Kuaishou Kling 1.5 SEP 2024 1129 950 900 ງ[[MiniMax Hailuo Al SEP 2024 1101 850 CogVideoX-5B • Pyramid Flow 800 **OpenAl's Sora** 腾计 Hunyuan Video DEC 2024 Tencent 1071 announced Feb '24, 750 Google's Veo 1 700 announced Mar '24 Mochi 1 OCT 2024 Genmo 1064 LTX Video 650 600 01/24 02/24 03/24 04/24 05/24 06/24 07/24 08/24 09/24 10/24 11/24 12/24 01/25

Video Generation Arena Leaderboard – Top 5 Models

Release

Arena

About the arena

Arena ELO Score¹

Model Release Date

The Artificial Analysis Video Generation Arena uses crowdsourced preferences to assess video generation model quality. As of December 2024, over 200k preferences have been gathered. These preferences are used to compute Arena ELO scores, which reflect the relative quality of video generation models.



TEXT TO SPEECH LANSCAPE

Speech Arena ELO vs. Model Release Date

Latest generation transformer-based Text to Speech models achieved new quality milestones in 2024, leapfrogging long-standing Hyperscaler offerings



Speech Arena Leaderboard – Top 5 Models

The Artificial Analysis Speech Arena uses crowdsourced preferences to assess text to speech model quality. As of December 2024, over 25k preferences have been gathered. These preferences are used to compute Arena ELO scores, which reflect the relative quality of text to speech models.

1. Artificial Analysis Text to Speech Arena ELO, calculated as of 16 December 2024



SPEECH TO TEXT LANDSCAPE

OpenAI reshaped AI transcription in late 2022 when it open-sourced Whisper, allowing cloud inference players to enter the market and compete on speed and price

Transcription Speed (Speed Factor)

Input audio seconds transcribed per second



Key takeaway

Cloud API offerings for OpenAI's open weights Whisper model now achieve Speed Factors well above 100 – the fastest of them can transcribe a full hour of audio in ~10 seconds

Transcription Pricing

USD per 1000 minutes of audio



Key takeaway

The price of transcription dropped to **under US\$1 per 1000 minutes of audio**. The latest Whisper endpoint release from Groq represents **a** ~72x **decrease** from Amazon's 2018 model and ~11x speed boost





Artificial Analysis

contact@artificialanalysis.ai

https://artificialanalysis.ai/

Legal notice: Copyright © 2024 Artificial Analysis, Inc. All rights reserved.

This document, including any data, analysis, and insights contained herein, is provided by Artificial Analysis for informational purposes only. The information is based on data collected through various sources, including but not limited to first party benchmarking and surveys conducted on our website. While Artificial Analysis strives to ensure the accuracy and reliability of the information, it is provided "as is" and may not be complete or up to date. The content should not be construed as professional advice, and recipients are encouraged to conduct their own research and analysis before making any decisions based on this information. By accessing or using this document, you agree to be bound by Artificial Analysis's Terms of Service, available on our website.