AI MODEL INTELLIGENCE INDEX: A DATA-DRIVEN BENCHMARKING REPORT

Apekshya Sharma

1. Introduction

Artificial intelligence has advanced significantly in recent years, with modern AI models now capable of handling complex tasks that once required human intellect. As these technologies become more deeply embedded in essential industries like healthcare, finance, and education, assessing their capabilities is more important than ever.

As artificial intelligence (AI) models continue to evolve, understanding and comparing their capabilities has become increasingly important for both researchers and industry professionals. This report aims to provide a structured analysis of leading AI models using data collected from the leaderboard on Artificial Analysis. The goal is to evaluate and compare model performance across a variety of key metrics.

The analysis focuses on several core areas:

1. Artificial Intelligence Index
2. Coding Index
3. Math Index
4. Relationships between Intelligence and Price, as well as Output Speed
5. Token Usage and Cost Efficiency
6. Context Window Size

Using Python and Jupyter Notebook, this report visualizes and interprets the collected data to provide a clearer understanding of how various AI models perform in different domains. The insights generated aim to guide users, developers, and decision-makers in selecting the most appropriate AI model based on their specific needs and priorities.

1.1 Background

The rapid development of artificial intelligence (AI) models has sparked significant interest in evaluating their capabilities across various domains. As new models emerge, it's important to assess not just their raw performance, but also how they compare efficiency, cost, and specialized skills such as coding and mathematics. With AI becoming an integral part of sectors such as education, software development, research, and business automation, benchmarking these models through reliable, standardized metrics is essential.

ArtificialAnalysis.ai is one such platform that provides a structured leaderboard comparing the capabilities of popular AI models. It evaluates models using indices such as the Intelligence Index, Coding Index, and Math Index, as well as metrics like price, output speed, token usage, and context window size. These benchmarks help users better understand the trade-offs between different models and make informed decisions.

1.2 Objectives of the report

This report aims to:

- Analyze and compare leading AI models based on data from ArtificialAnalysis.ai.
- Evaluate performance across the Artificial Intelligence Index, Coding Index, and Math Index.
- Explore the relationships between Intelligence vs. Price and Intelligence vs. Output Speed.
- Examine Token Usage and Cost Efficiency of different models.
- Compare models based on Context Window size to assess their ability to handle larger inputs.
- Present visual insights that highlight key trends, strengths, and limitations of each model.

The goal is to offer a clear, data-driven understanding of how AI models perform across multiple dimensions, enabling developers, researchers, and decision-makers to choose the most suitable model for their specific needs.

1.3 Data Source and Tools Used

All data for this analysis was sourced from the public leaderboard available on

ArtificialAnalysis.ai, a platform that ranks AI models using consistent and transparent evaluation metrics.

To analyze and visualize the data:

- Jupyter Notebook was used as the development environment
- The analysis was performed using Python, with libraries such as:

i. Pandas for data handling, ii. Matplotlib
and Seaborn for visualizations, iii. NumPy for numerical
operations.

This combination of tools enabled efficient data manipulation, clear charting, and meaningful interpretation of results.

2 Methodology

2.1 Data Collection

For this study, the data was sourced directly from the leaderboard hosted on ArtificialAnalysis.ai. This site regularly updates performance statistics for a variety of AI models, offering a consistent benchmark for comparison. It includes detailed scores that reflect model capabilities in different index areas such as reasoning, code generation, math, speed, and pricing.

2.2 Metrics Selected for Analysis

To carry out a thorough comparison of the AI models, the following key metrics were selected:

1. **General Intelligence Score** – Indicates how well a model performs across tasks involving logic, reasoning, and language.
2. **Coding Ability Score** – Measures the effectiveness of the model in solving programmingrelated tasks.
3. **Mathematical Proficiency Score** – Reflects the model's performance in solving math problems.
4. **Performance vs. Cost** – Explores how the model's intelligence score relates to its usage cost.
5. **Performance vs. Output Speed** – Examines the balance between model intelligence and how quickly it generates responses.
6. **Token Usage and Cost** – Assesses how efficiently a model uses tokens in relation to its pricing.

7. **Context Window Size** – Refers to the number of tokens a model can handle in one input, which impacts its ability to manage longer prompts or documents.

These evaluation points were selected to ensure a balanced review of each model's capabilities and practicality.

3 Analysis and Results

3.1 Overview of the model performance

The dataset compiled from ArtificialAnalysis.ai contains a range of leading AI models from developers such as OpenAI, Google, Anthropic, xAI, and others. Each model is evaluated across multiple dimensions, including general intelligence, coding ability, math reasoning, input/output costs, processing speed, and context window size.

A quick look at the dataset shows that:

- OpenAI’s GPT-5 series (high, medium, and low variants) consistently achieve the highest scores in the Artificial Intelligence Index, ranging between 63–69. These models also demonstrate strong coding and math abilities, particularly in the high and medium versions.
- xAI’s Grok 4 and Google’s Gemini 2.5 Pro also perform competitively, with Grok 4 scoring the highest in coding ability (88).
- Cost efficiency varies widely: smaller variants like GPT-5 mini and GPT-5 nano achieve significantly higher Intelligence per USD, highlighting their value despite lower raw intelligence scores. For instance, GPT-5 nano delivers the best cost-effectiveness ratio with an Intelligence per USD of 1080.0, far exceeding larger models.
- Anthropic’s Claude models show decent reasoning scores but come with considerably higher costs (e.g., Claude 4.1 Opus Thinking with $15 input cost per 1M tokens and $75 for output).
- Models with extended context windows (up to 1M tokens, e.g., Gemini 2.5 Flash and Grok 3 mini) are designed for large-scale tasks, though their intelligence scores fall slightly below the highest-ranking GPT-5 models.

- Speed efficiency (tokens per second per dollar) is highest for smaller and optimized models such as GPT-5 nano and Grok 3 mini-Reasoning (high), showing that lightweight models can achieve extremely high throughput at minimal cost.

Overall, the data set reflects a clear trade-off:

1. High-end models dominate in raw intelligence and reasoning but come with high cost per token.
2. Mid-range and lightweight models achieve exceptional efficiency in terms of speed and cost, making them suitable for more practical, large-scale deployments.

3.2 Artificial Analysis Intelligence Index

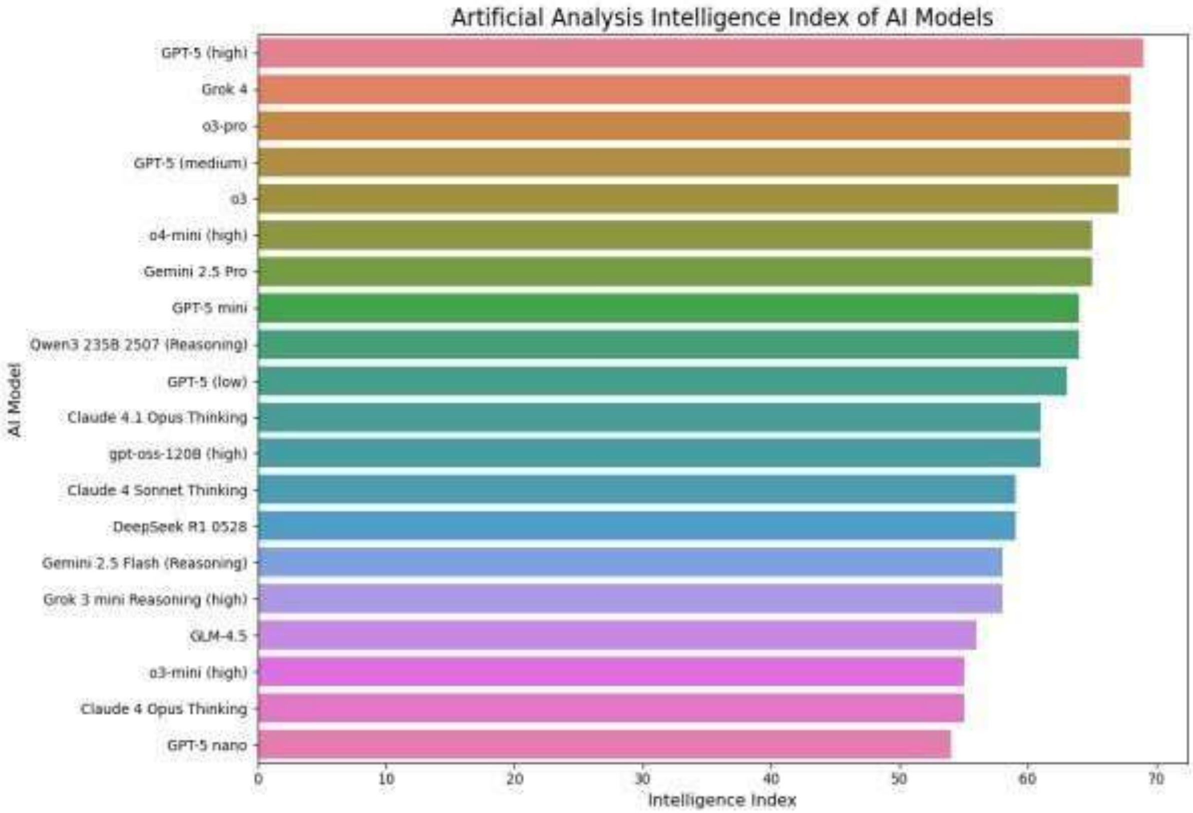


Figure 1 Artificial Analysis Intelligence Index

The Artificial Analysis Intelligence Index measures the overall reasoning and problem-solving capability of AI models. The visualization shows that OpenAI's GPT-5 (high) stands out as the top performer with the highest intelligence score, followed closely by Grok 4, o3-pro, and GPT-5 (medium). These models consistently score near the top, reflecting their advanced reasoning and general-purpose abilities.

i. GPT-5 (high) ranks the highest in intelligence, followed closely by Grok 4, o3-pro, and GPT-5 (medium).

ii. o4-mini (high) and Gemini 2.5 Pro deliver competitive reasoning performance despite being optimized or smaller models.

iii. Claude series (Anthropic) and Gemini 2.5 Flash (Google) fall in the mid-range (scores ~55–60), slightly below OpenAI's flagship models. iv. Lightweight variants like GPT-5 nano and o3-mini (high) score lower (mid-50s), showing efficiency-focused models trade off reasoning depth.

v. Overall trend: larger flagship models dominate in reasoning, while smaller models prioritize efficiency over raw intelligence.

3.3 Artificial Analysis Math Index

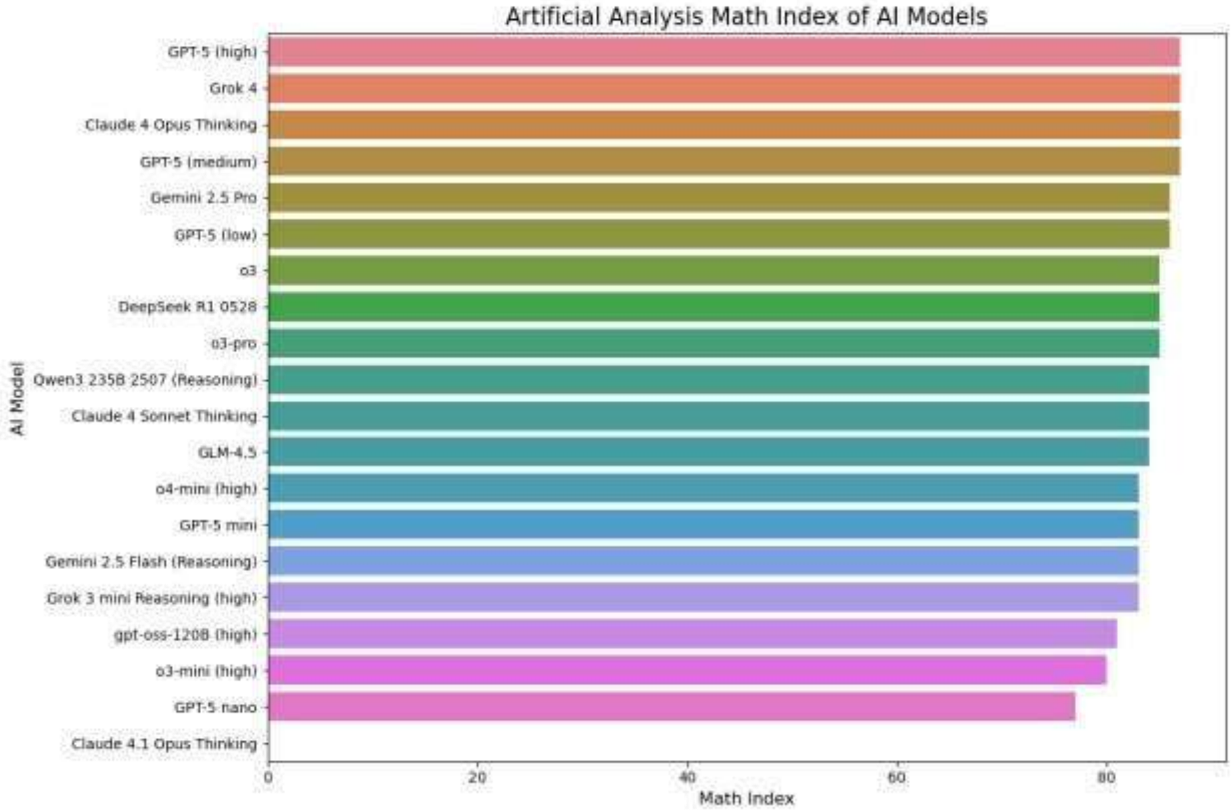


Figure 2 Artificial Analysis Math Index

The Math Index evaluates AI models on their ability to solve complex mathematical problems, ranging from basic numerical reasoning to advanced problem-solving tasks. Performance in this domain reflects a model's logical consistency, accuracy in computation, and capacity to handle abstract reasoning. Models with higher scores demonstrate stronger analytical and quantitative abilities, making them more reliable for math-intensive applications.

Top performers in math reasoning are:

i. GPT-5 (high), Grok 4, Claude 4 Opus Thinking, and GPT-5 (medium) – all scoring at the very top.

ii. Gemini 2.5 Pro and GPT-5 (low) also perform strongly, very close to the leaders.

iii. Mid-performing models include o3, DeepSeek R1 0528, o3-pro, and Qwen3 235B, with slightly lower but still competitive math capabilities.

iv. Claude 4 Sonnet Thinking, GLM-4.5, and o4-mini (high) remain in the middle range.

v. Smaller/optimized variants like GPT-5 mini, Gemini 2.5 Flash, Grok 3 mini-Reasoning, and o3-mini (high) score moderately lower, showing some trade-off in mathematical depth for efficiency.

vi. Lowest scores are seen in GPT-5 nano and Claude 4.1 Opus Thinking, indicating these models are less optimized for advanced mathematical reasoning. vii. Overall trend: Larger flagship models dominate in math reasoning, while smaller/minimized versions sacrifice performance for cost or efficiency.

3.4 Artificial Analysis Coding Index

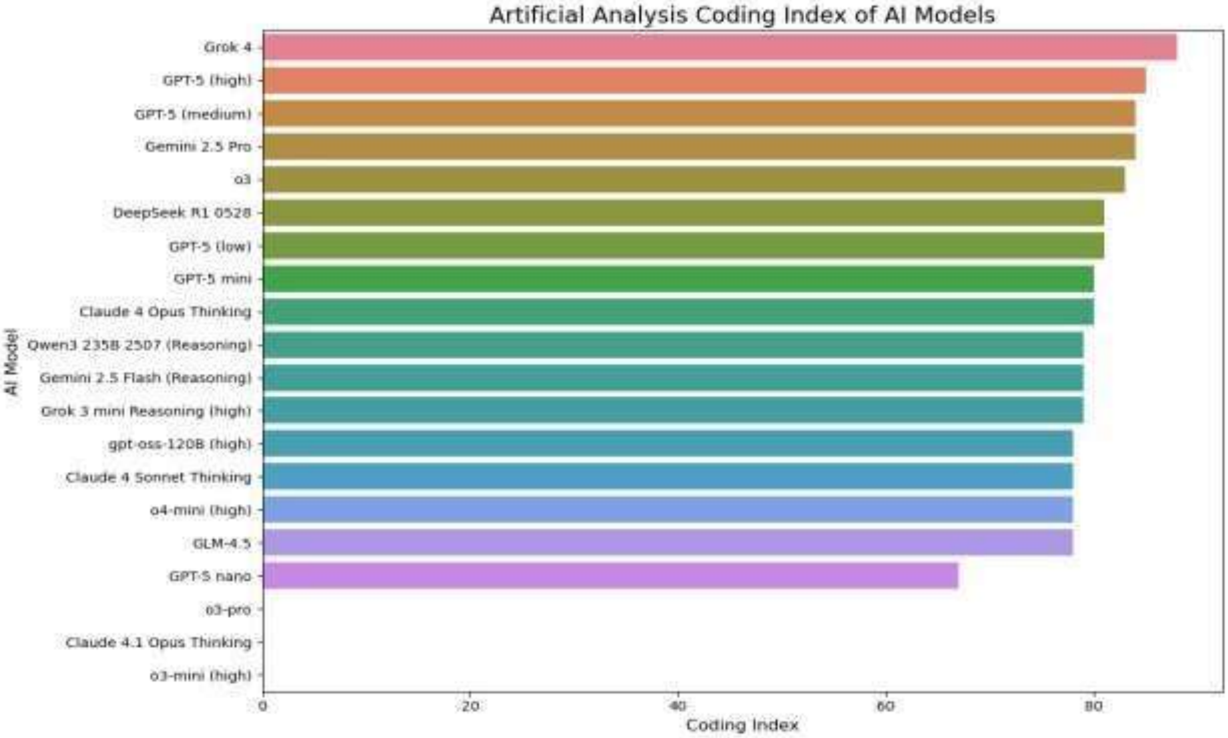


Figure 3 Artificial Analysis Coding Index

This chart presents the Artificial Analysis Coding Index of various leading AI models, comparing their relative performance in coding-related tasks. The index provides a numerical representation of how effectively each model handles programming, reasoning, and technical problem-solving. By ranking these models, we can observe trends in performance across different AI families such as GPT, Claude, Gemini, Grok, O-series, and others.

A. Top Performers:

- Grok 4 leads with the highest coding index, slightly ahead of GPT-5 (high) and GPT-5 (medium), indicating superior performance in complex coding and reasoning tasks.
- Gemini 2.5 Pro and o3 also rank very close, showing competitive capabilities. B. Strong Mid-Tier Models:

- Models like DeepSeek R1 0528, Claude 4 Opus Thinking, and Qwen3 235B (Reasoning) cluster in the middle range, demonstrating consistent reliability though slightly below the top-tier performers.
- GPT-5 (low) and GPT-5 mini still perform well, proving that even scaled-down versions retain strong coding proficiency. C. Lower End of the Spectrum:
- GPT-5 nano, o3-pro, and o3-mini (high) show significantly lower coding index scores compared to others, reflecting their lightweight or specialized nature, likely optimized for efficiency over raw problem-solving power. D. General Trend:
- Larger, more advanced models (e.g., GPT-5 high/medium, Grok 4, Gemini 2.5 Pro) dominate the upper tier.
- Smaller or efficiency-focused models trade off raw coding performance for speed, cost, or specialized reasoning.

3.5 Intelligence per USD (Value for Money)

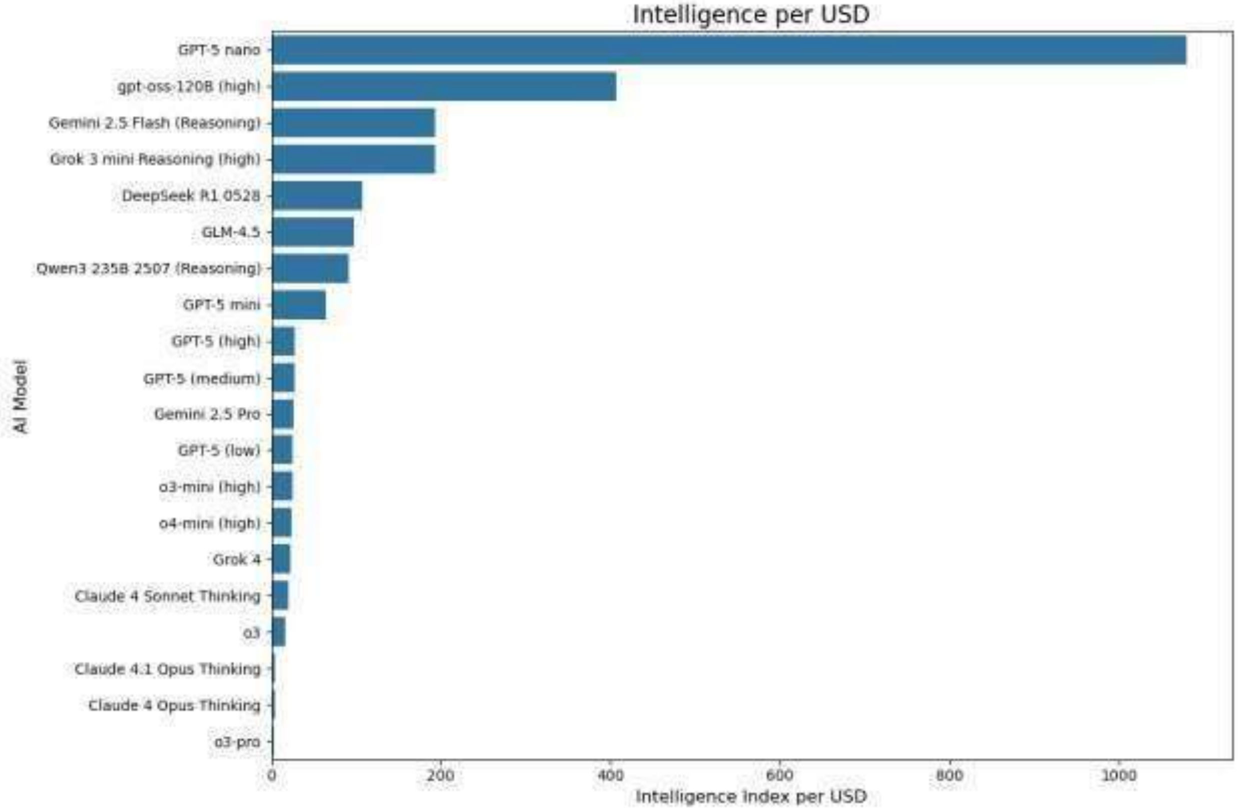


Figure 4 Intelligence per USD

i. Best value: GPT-5 nano gives the most intelligence for each dollar, very costeffective.

ii. Good value: gpt-oss-120B, Gemini 2.5 Flash, and Grok 3 mini-Reasoning also offer strong performance for the price.

iii. Middle ground: Models like DeepSeek R1, GLM-4.5, and Qwen3 give okay value but not outstanding.

iv. Low value: Big premium models (Claude 4 Opus, Claude 4.1, o3-pro) are powerful but very expensive, so they don't give much intelligence per dollar.

Intelligence Index vs. Price Index (Average token cost)

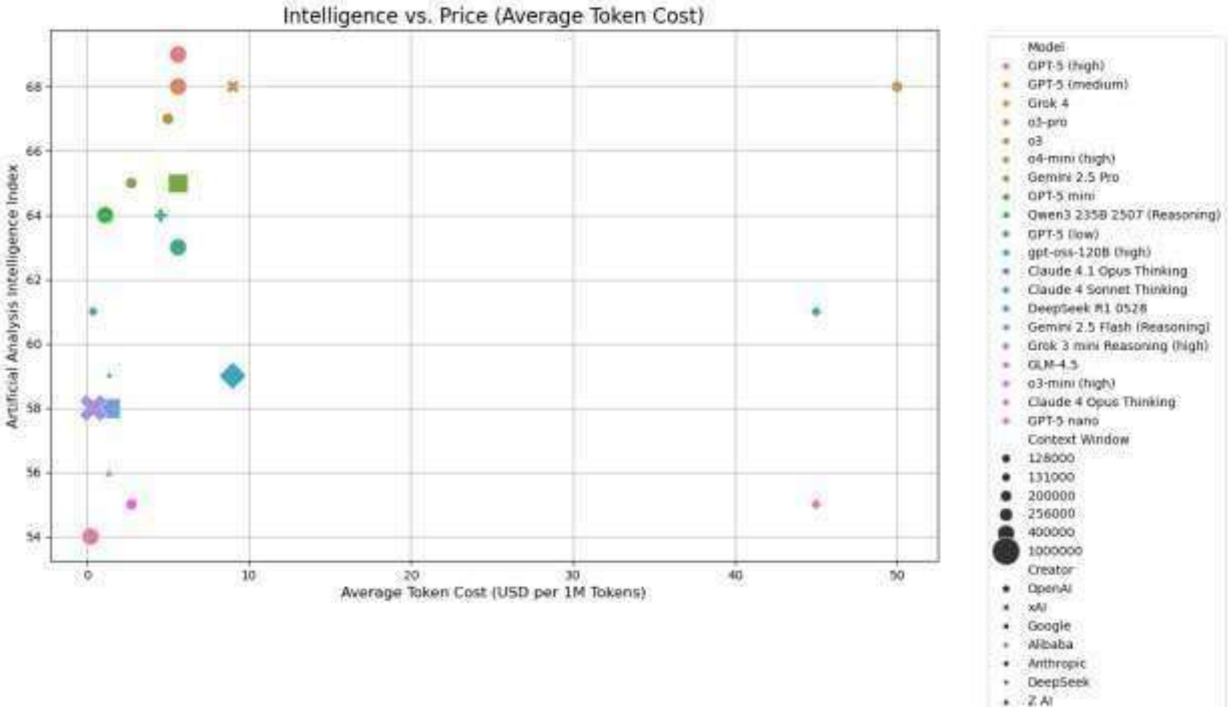


Figure 5 Intelligence vs. Price per token

i. AI models range in intelligence from 54 to 68 on the Artificial Intelligence Index.

ii. Costs vary from nearly 0 to over 50 USD per 1M tokens.

iii. High-intelligence models like GPT-5 (high) and Claude 4.1 Opus (around 68) are affordable, costing under 20 USD.

iv. Low-cost models like GPT-5 nano (near 0 USD) offer moderate intelligence (around 54).

v. Expensive models like Creator (over 40 USD) don't always have the highest intelligence.

vi. xAI and OpenAI models span a wide range of intelligence and cost levels. vii. Some providers, like DeepSeek and Z AI, focus on lower intelligence models.

Intelligence Index vs. Output Speed

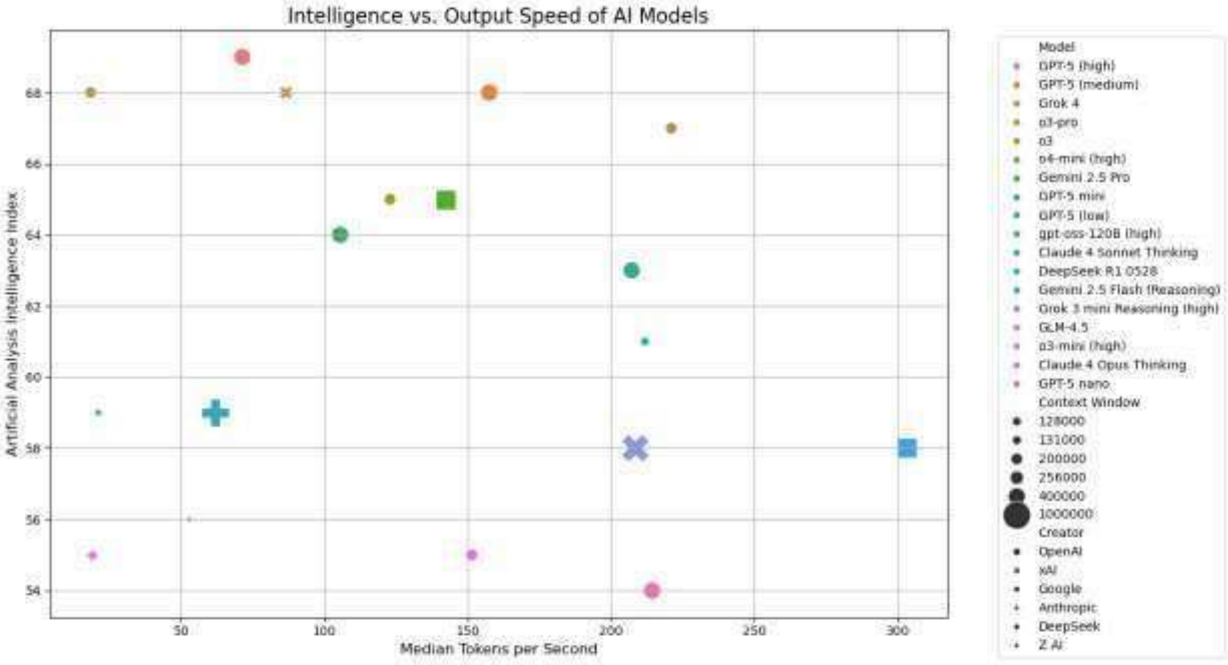


Figure 6 Intelligence Index vs. Output speed

i. AI models range in intelligence from 54 to 68 on the Artificial Intelligence Index.

ii. Output speed varies from 50 to over 300 median tokens per second.

iii. High-intelligence models like GPT-5 (high) and Grok 4 (around 68) achieve speeds from 100 to 200 tokens per second.

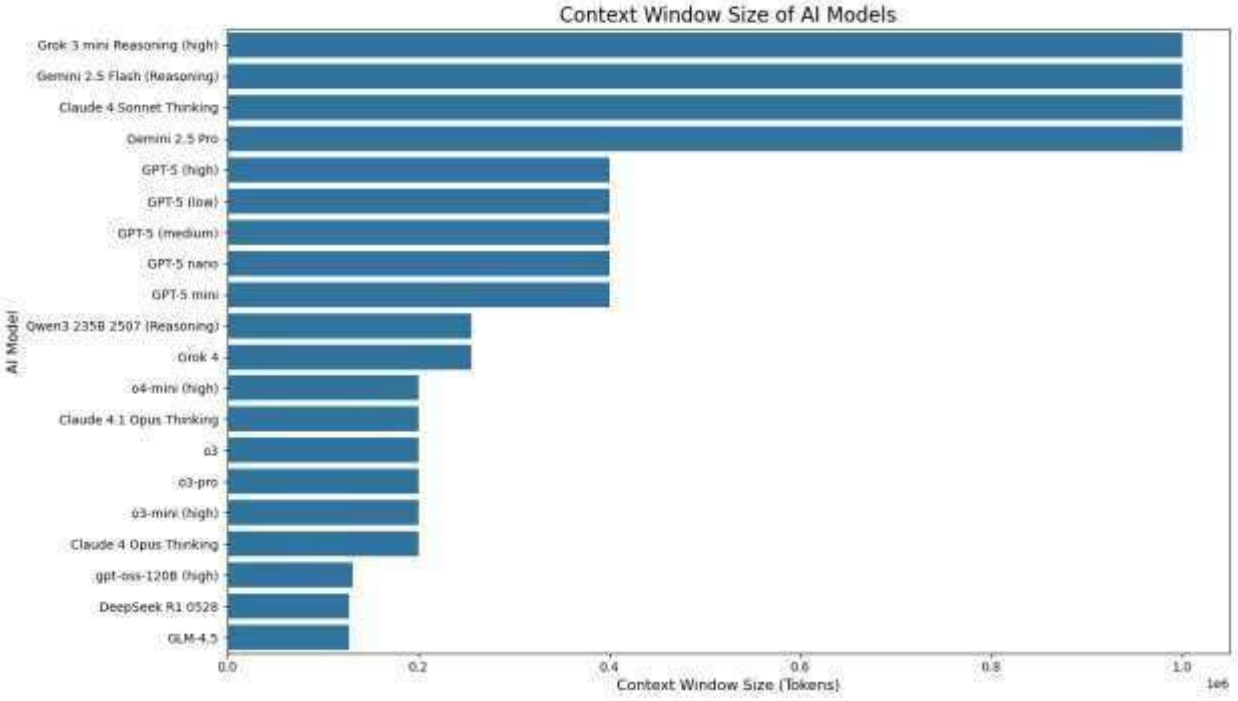
iv. Low-intelligence models like GPT-5 nano (around 54) have the slowest speeds (below 50 tokens per second).

v. Models with moderate intelligence (60-64) show a wide speed range, from 50 to over 250 tokens per second.

vi. xAI and OpenAI models are distributed across both high intelligence and varying speeds.

vii. Some models (e.g., Creator, 4000000 context) with higher speeds (around 250-300) don’t always have top intelligence.

4 Context Window



The bar chart displays the context window size (in millions of tokens) for various AI models.

i. Models like Grok 3 mini Reasoning (high), Gemini 2.5 Flash (Reasoning), and Claude 4 Sonnet Thinking have the largest context windows, exceeding 1 million tokens.

ii. Mid-range context windows (around 0.4-0.6 million tokens) are common among models like GPT-5 (high), GPT-5 (low), and GPT-5 (medium).

iii. Smaller context windows (below 0.2 million tokens) are seen in models like DeepSeek R1 0528 and GLM-4.5.

iv. xAI models (e.g., Grok 3, Grok 4) and OpenAI models (e.g., GPT-5 variants) show a wide range of context window sizes, from small to very large.

v. Some models, such as Claude 4 Opus Thinking and o3-mini (high), have moderate context windows (around 0.2-0.3 million tokens).