Datasaur

Foundation Model Evaluation Report

2024 Report

01 Introduction

Artificial Intelligence (AI) is often perceived as a monolithic entity – a singular product to be judged as either effective or ineffective. However, the reality is far more nuanced. The diversity of models, and their potential for future specialization, illustrates that AI is not a one-size-fits-all solution. The key differences when evaluating AI models aren't simply proprietary versus open-source; rather, they are rooted in three primary dimensions: cost, speed, and quality. The interplay of these factors determines whether an AI project is cost-effective, timely, and impactful.

To assess models effectively, evaluation must occur on a per-inference basis. This approach provides the most universal and equitable measure, as different models process varying amounts of tokens per request, scale from proof of concept to full production with varying efficiency, and deliver different quality levels that can be scored individually.

Additionally, we've utilized broad categories of model requests to offer insights into which models perform best in specific scenarios. These categories include content generation, customer support, and others, ensuring that our evaluations reflect the most common and practical use cases. To further refine our assessments, we incorporated benchmark datasets such as MMLU, IFEval, Fin-Fact, CareQA and MATH, which allow us to gauge each model's performance across diverse tasks, from general knowledge and instruction-following to complex mathematical reasoning.

02 General Thoughts

The spectrum of costs, speeds, and result quality across AI models is vast, with one consistent theme: a lack of uniformity. This diversity is beneficial, as it underscores that no single model excels at every task. While some models may consistently deliver the highest accuracy, they often come with the trade-off of higher costs or slower inference speeds. The variety also extends to significant discrepancies – some models are priced more than 70% lower than others, or operate at half the speed, while others may produce results that are entirely unusable.

These differences highlight the importance of testing multiple models for each project to identify the optimal return on investment (ROI) rather than relying on just one. It's crucial to balance the project's needs with the unique combination of cost, speed, and quality.

For instance, you might opt to:

- Accept a 60% reduction in cost for a minimal drop in accuracy, suitable for non-critical tasks.
- Choose a model with a slower inference speed but superior quality for tasks that require high precision, such as legal document analysis.
- Prioritize a faster, slightly less accurate model when speed is essential, like in real-time customer support applications.

This variability is why we developed this report. While it provides comprehensive insights into the best models to consider for your AI project, nothing can replace the value of sandboxing your unique needs across multiple models to find the best fit.

03 Benchmark Datasets Overview

We select five diverse datasets as benchmarks to conduct a comprehensive evaluation, helping you choose the right model for your use case across three key dimensions: speed, cost, and quality. These datasets were carefully chosen to represent real-world challenges, ensuring that your findings are more objective and that you can identify the optimal balance for your specific goals

Massive Multitask Language Understanding

The MMLU (Massive Multitask Language Understanding) dataset is a comprehensive benchmark designed to evaluate the performance of language models across a broad spectrum of tasks and domains. Notable for its scale, diversity, and task difficulty, MMLU includes 57 different subjects spanning various fields such as humanities, STEM, social sciences, and more. This makes it one of the largest and most diverse benchmarks available for assessing language models. The tasks range from elementary-level questions to advanced, university-level problems, requiring different types of reasoning, such as factual recall, mathematical problem-solving, reading comprehension, and logical analysis.

Instruction-following Capabilities of LLM

The IFEval dataset, introduced by Jeffrey Zhou and colleagues, is specifically designed to evaluate the instruction-following capabilities of large language models (LLMs).

The dataset comprises a collection of prompts, each associated with verifiable instructions that a language model must follow accurately. This focus on instruction-following is crucial, as it directly impacts the real-world applicability of LLMs in various tasks. To assess the accuracy of instruction-following, IFEval employs several metrics. These include prompt-level strict-accuracy, which measures the percentage of prompts where all instructions are followed correctly, and inst-level strict-accuracy, which evaluates the accuracy at the level of individual instructions. Additionally, there are looser accuracy metrics that provide a more lenient evaluation, both at the prompt and instruction levels.

Comprehensive Mathematical Problems Solving

The MATH dataset, created by Dan Hendrycks and collaborators, is a comprehensive collection of mathematical problems designed to evaluate the problem-solving abilities of artificial intelligence systems. This dataset encompasses a wide range of topics typically covered in high school mathematics curricula. These topics include algebra, geometry, calculus, statistics, and more, ensuring that the dataset provides a robust challenge to AI models. Each problem in the dataset is accompanied by a detailed solution, which not only includes the final answer but also the step-by-step reasoning required to arrive at that answer. This feature is crucial for training and evaluating AI systems, as it allows for a more nuanced assessment of their problem-solving processes.

The MATH dataset includes problems that require a deep understanding of mathematical concepts and the ability to apply these concepts in novel ways. The dataset helps to push the boundaries of what AI can achieve in the realm of mathematics for both AI testing and educational purposes. This MATH dataset also encourages the development of models that can provide transparent and understandable reasoning, which is crucial for applications in education, science, and engineering.

Financial Knowledge and Fact Checking

Fin-Fact is a newly developed benchmark dataset specifically designed for multimodal financial fact-checking and explanation generation. It addresses the significant issue of misinformation in the financial sector, which can adversely affect public trust and investor decisions. The dataset comprises 3,369 claims related to various financial topics, including economy, taxes, and debt. Each claim is categorized as True, False, or Not Enough Information (NEI) based on expert annotations. This structured approach allows for a comprehensive analysis of the factuality of claims, making it a valuable resource for researchers and practitioners in the field.

Additionally, each claim in the dataset is accompanied by detailed justifications from professional fact-checkers. This enhances the dataset's credibility and allows for the generation of transparent explanations. By providing insights into the reasoning behind fact-checking decisions, Fin-Fact empowers users to understand the complexities of financial claims, which is essential for fostering trust in financial reporting.

Healthcare Question Answering

The CareQA dataset is a medical question-answering dataset, designed for multiple-choice QA tasks. It is based on exams for the Spanish Specialized Healthcare Training (FSE) and spans subjects such as medicine, biology, chemistry, nursing, pharmacology, and psychology. The dataset covers questions from 2020 to 2024, available in both English and Spanish. Originally sourced from Spanish exams, the English versions were translated using GPT-4, with some manual review for quality assurance.

CareQA contains 5,621 samples, where each entry presents a question with four possible answers, and the goal is for models to select the correct one. It's widely used for evaluating model accuracy in multiple-choice scenarios, especially in healthcare-related fields

04 Foundation Models Overview

We select six popular foundation models, including both open source and proprietary options, for evaluation. We assess their performance across multiple benchmarks to analyze the strengths and weaknesses of each model for specific tasks.

Claude 3.5 Sonnet

Anthropic's Claude 3.5 Sonnet is a language model distinguished by its creative flair. It excels at generating diverse text formats, from poems and scripts to detailed narratives. This model is particularly adept at tasks demanding nuanced understanding and expression.

Gemini 1.5 Flash

Google's Gemini 1.5 Flash prioritizes speed and efficiency without compromising performance. Designed for swift response times, it's ideal for applications requiring immediate outputs, such as real-time translation, summarization, or answering queries.

GPT-4o

OpenAI's GPT-4o is a versatile powerhouse known for its human-like text generation capabilities. It handles a vast array of tasks, including translation, creative writing, answering complex questions, and providing comprehensive summaries.

GPT-4o Mini

A scaled-down version of GPT-4o, GPT-4o Mini offers a balance between performance and efficiency. While retaining many of the original model's strengths, it's well-suited for resource-constrained environments or applications demanding quicker response times.

Llama 3.1 8B Instruct

Meta AI's Llama 3.1 8B Instruct is an open-source model specialized in following instructions. It effectively generates text formats that adhere to given prompts or guidelines, making it a valuable tool for tasks requiring specific outputs.

Mistral 7B Instruct v0.2

Another open-source model, Mistral 7B Instruct v0.2 is designed to excel at instruction following. Despite its smaller size, it demonstrates competence in producing text according to given prompts or commands.

05 Benchmarking Results and Analysis

This section will analyze the top six models in five different benchmark datasets to provide insight to the best performing models across three critical categories, Quality, Cost, and Speed. While there is no definitive answer to what is the “best” model there are models that perform better against different benchmarks. Additionally each AI project, regardless of similarity or adjacency to a specific benchmark, can have unique characteristics that warrant specific tests across the same criteria referenced above.

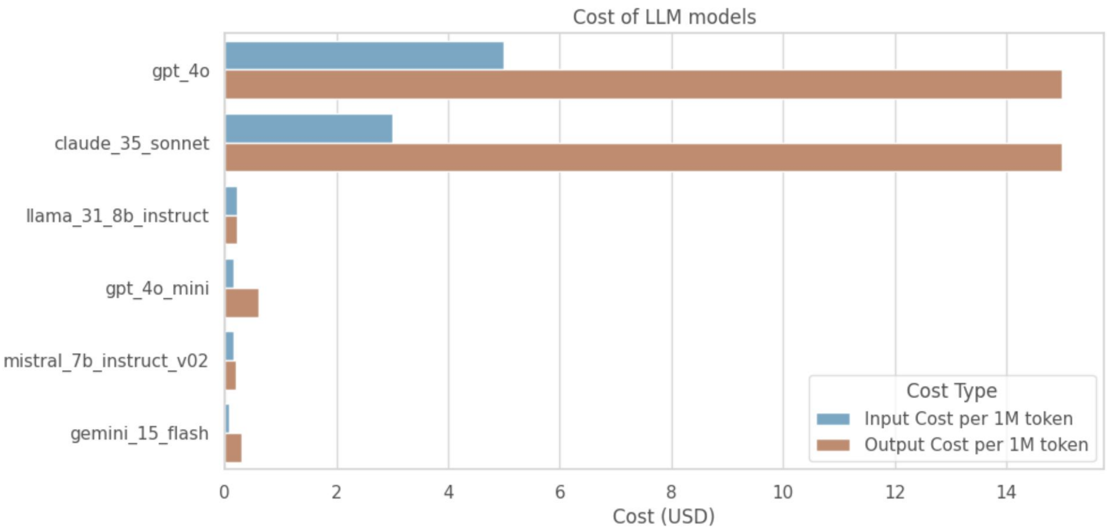
This data is meant to be a directional guide for prioritization of models for specific projects. We always encourage organizations to test models against their specific use case using a platform like LLM Labs or a similar service.

Additionally your project might have different prioritizations. For example, in use cases like chatbots, speed and cost may be crucial factors. In contrast, for sensitive applications such as medical diagnostics, speed and cost can be deprioritized in favor of achieving high-quality, accurate results.

Cost of LLM Models

Cost is a crucial factor when selecting an LLM for specific use cases. While one model might excel in both speed and quality, there are scenarios where we need to prioritize low cost, even if it means sacrificing performance and speed. This is particularly true for non-critical tasks or simpler processes where high precision or fast response times are less essential.

When it comes to cost, we need to consider both input and output costs. Input cost refers to the expense of processing a prompt, question, or instruction with the LLM. On the other hand, output cost relates to the expense of generating the response. We do have control over these costs – input costs can be optimized by crafting concise prompts, and output costs can be managed by setting a maximum response length.



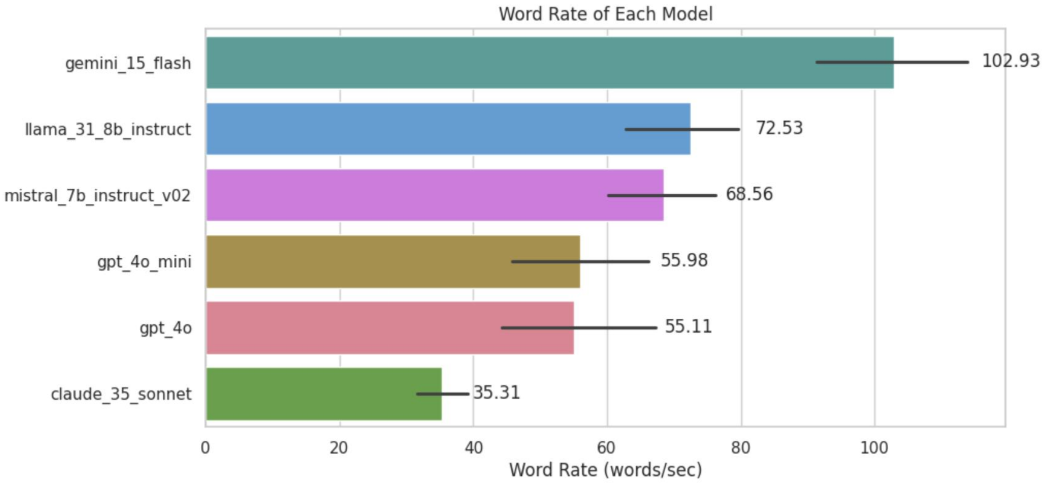
Two major models, GPT-4o and Claude 3.5 Sonnet, both proprietary, are significantly more expensive than other models. This higher cost likely reflects the advanced features they offer, particularly in achieving higher precision in results.

For less critical tasks where precision is not as crucial, other proprietary models like Gemini 1.5 Flash and GPT-4.o mini offer more affordable options, making them competitive with open-source models such as Mistral 7B Instruct and Llama 3.1 8B Instruct.

Open-source models provide more flexibility, especially when customization is needed for specific use cases. While fine-tuning these models can be costly at first, continuous improvement allows for better results over time, while still maintaining control over costs and processing speed.

LLM Speed

Speed is another crucial factor to consider when selecting a model, especially for real-time applications like chatbots or AI assistants that require rapid responses. LLMs can vary significantly in their processing speed due to several factors, including available computing resources, model architecture, and input length. It's important to note that speed is relative to the specific use case. For applications where quick response times are essential, this factor should be given greater weight in the model selection process.



Overall, Gemini 1.5 Flash stands out as the fastest model, with a significant gap compared to other LLMs, especially when compared to Claude 3.5 Sonnet, which is nearly three times slower. This is expected, as Gemini 1.5 is designed for users who prioritize fast response times without compromising performance. Additionally, the two open-source models, Llama 3.1 8B and Mistral 7B, also demonstrate impressive performance, slightly outpacing both GPT-4o and GPT-4o Mini.

LLM Quality

The quality of an LLM can be measured by assessing how closely the model's answers align with the source of truth. The more precise the answers, the higher the quality of the model. In this section, we will present the quality of each LLM across various benchmarks, allowing us to identify similar use cases and determine which model is most suitable based on three criteria: quality, speed, and cost.

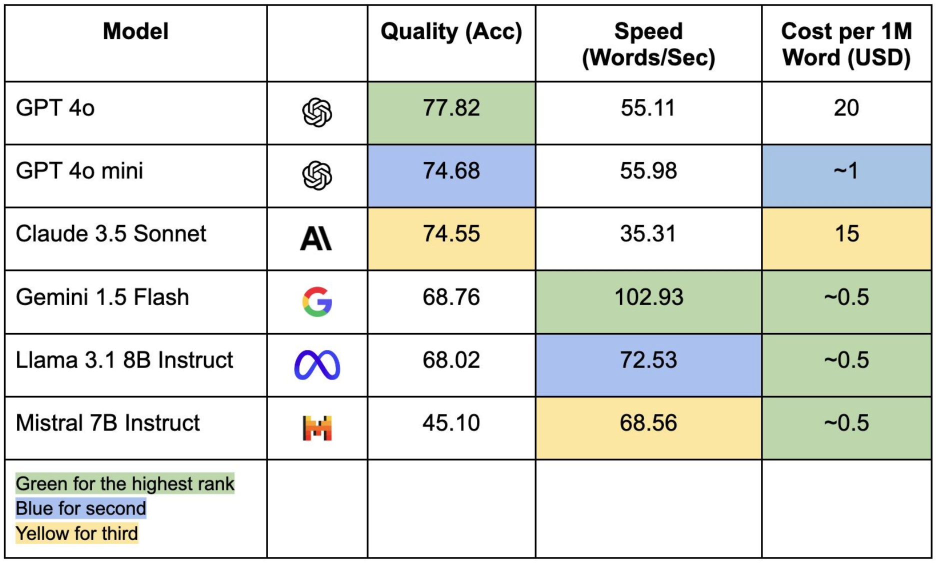
MMLU Benchmark Result (Broad Spectrum Dataset)

Claude 3.5 Sonnet achieves the highest accuracy score, closely followed by GPT 4o with a marginal difference of only 0.53%. Notably, proprietary models occupy the top three positions in terms of quality. For use cases similar to the MMLU dataset where quality of results is a priority, but speed and cost are also important considerations, Claude 3.5 Sonnet emerges as a strong option. It offers a more cost-effective solution compared to GPT 4o, albeit with a slightly longer processing time. However, if cost and speed are the primary concerns while still maintaining relatively high quality output, GPT 4o mini presents an attractive alternative. It delivers high-quality results comparable to the top models, matches them in speed, and ranks among the three most affordable options in the comparison.

Model		Quality (Acc)	Speed (Words/Sec)	Cost per 1M Word (USD)
Claude 3.5 Sonnet	A\	84.27	35.31	15
GPT 4o		83.74	55.11	20
GPT 4o mini		76.43	55.98	~1
Mistral 7B Instruct	M	56.18	68.56	~0.5
Gemini 1.5 Flash	G	45.15	102.93	~0.5
Llama 3.1 8B Instruct	∞	36.16	72.53	~0.5
Green for the highest rank Blue for second Yellow for third				

IFEval Benchmark Result (Instruction-Following Capabilities)

The proprietary models dominate the top four positions in this benchmark, with the GPT series securing first and second place, followed closely by Claude 3.5 Sonnet. GPT-4o achieved the best results, with a 3.14% lead over second place and a significantly larger margin – up to 10% – compared to open-source models like Llama 3.1 8B. The gap is even more pronounced with Mistral 7B, reaching up to 32%. GPT-4o is an excellent choice when quality is a crucial factor in your use case. However, if you can tolerate a slight drop in accuracy (around 3%) and prioritize speed and cost, GPT-4o Mini is the best option.



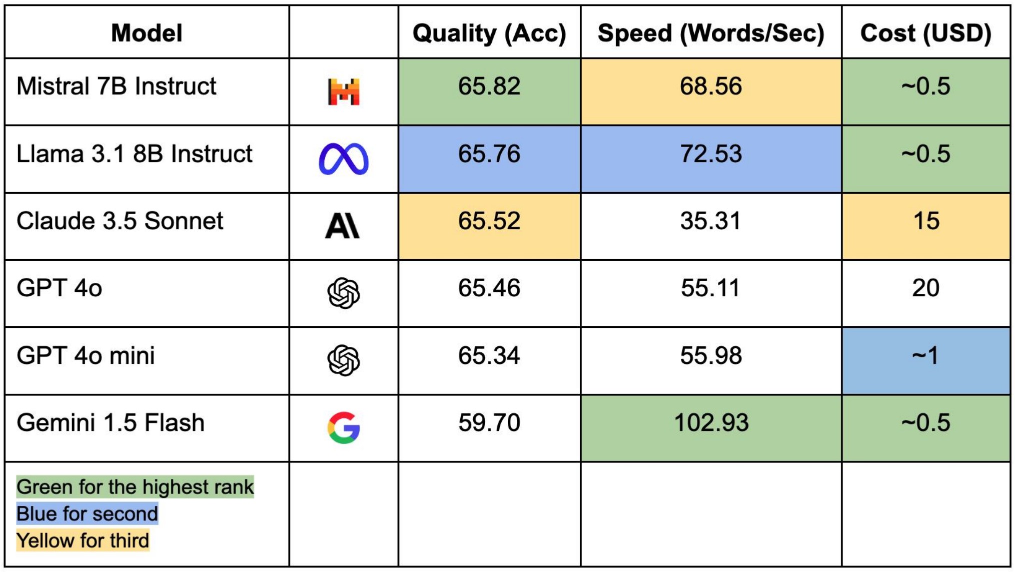
Model		Quality (Acc)	Speed (Words/Sec)	Cost per 1M Word (USD)
GPT 4o		77.82	55.11	20
GPT 4o mini		74.68	55.98	~1
Claude 3.5 Sonnet		74.55	35.31	15
Gemini 1.5 Flash		68.76	102.93	~0.5
Llama 3.1 8B Instruct		68.02	72.53	~0.5
Mistral 7B Instruct		45.10	68.56	~0.5
Green for the highest rank Blue for second Yellow for third				

MATH Benchmark Result (Problem-Solving Tasks)

The performance of models in problem-solving and complex tasks requiring explainability can be assessed using this benchmark. It's evident that all models struggle with mathematical tasks,   as none manage to achieve even 30% accuracy. Three proprietary models dominate the top four positions with minimal differences in their scores. Open-source models, like Mistral, were unable to solve the math tasks at all, while Llama 3.1 8B reached only 3.51% accuracy, which falls far short of expectations. Selecting a model for this task is challenging based on the three criteria (quality, speed, cost). However, as a user, I prefer to have more control over the model. Therefore, I would opt for an open-source model and attempt to improve its performance through fine-tuning for this specific use case.

Fin-Fact Benchmark Result (Financial Benchmark)

Financial use cases pose a challenging task, yet open-source models achieved the highest scores, surpassing both Claude 3.5 Sonnet and GPT-4o. This is a promising sign that open-source models can handle complex financial tasks, offering more flexibility for performance improvements through fine-tuning, while still delivering the best speed and lower costs. For users with similar use cases, open-source models like Mistral 7B Instruct and Llama 3.1 8B Instruct could be the optimal solution, covering all three key criteria: quality, speed, and cost.

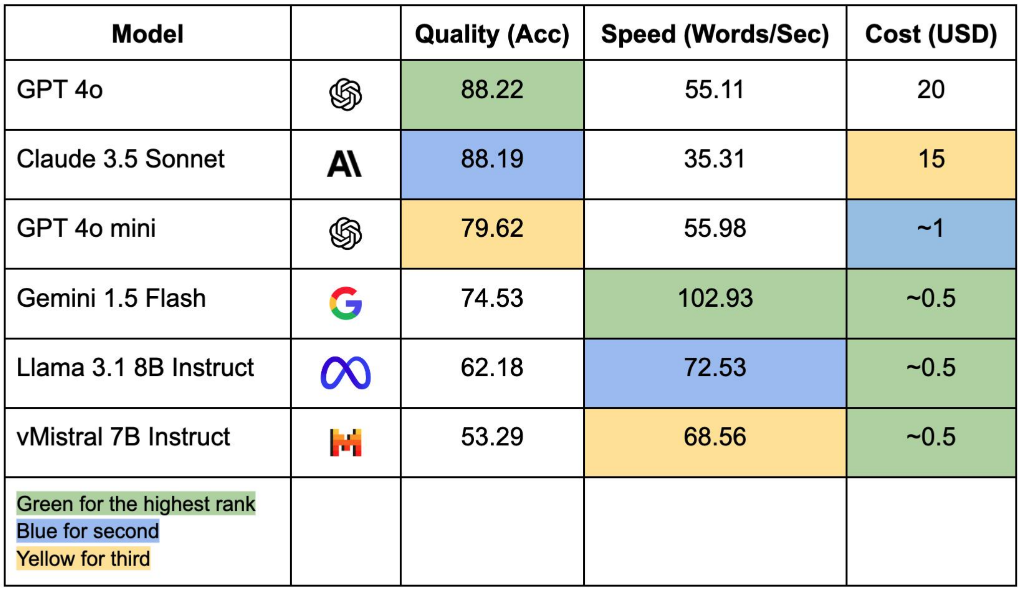


Model		Quality (Acc)	Speed (Words/Sec)	Cost (USD)
Mistral 7B Instruct		🟩 65.82	🟨 68.56	🟩 ~0.5
Llama 3.1 8B Instruct		🟦 65.76	🟦 72.53	🟩 ~0.5
Claude 3.5 Sonnet		🟨 65.52	35.31	🟨 15
GPT 4o		65.46	55.11	20
GPT 4o mini		65.34	55.98	🟦 ~1
Gemini 1.5 Flash		59.70	🟩 102.93	🟩 ~0.5
🟩 Green for the highest rank 🟦 Blue for second 🟨 Yellow for third				

CareQA Benchmark Result (Medical Dataset)

The proprietary models dominate the top four positions with a significant gap, up to ~35%, compared to open-source models. The health dataset is typically very sensitive and requires highly precise results, making it essential to prioritize quality over speed and cost. GPT-4o and Claude 3.5 Sonnet perform similarly, and both are worth considering for integration. Each excels in either speed or cost, so it ultimately depends on the user's priorities. If an accuracy drop of ~9% is acceptable, GPT-4o mini could be the best option for balancing quality, speed, and cost.

For even faster performance, Gemini 1.5 Flash stands out, being twice as fast as GPT-4o and GPT-4o mini, and three times faster than Claude 3.5 Sonnet. Despite a ~5% reduction in accuracy compared to GPT-4o mini, Gemini 1.5 Flash is worth considering for its cost efficiency – it is twice as cheap as GPT-4o mini and 30 to 40 times cheaper than GPT-4o and Claude 3.5 Sonnet.



Model		Quality (Acc)	Speed (Words/Sec)	Cost (USD)
GPT 4o		🟩 88.22	55.11	20
Claude 3.5 Sonnet		🟦 88.19	35.31	🟨 15
GPT 4o mini		🟨 79.62	55.98	🟦 ~1
Gemini 1.5 Flash		74.53	🟩 102.93	🟩 ~0.5
Llama 3.1 8B Instruct		62.18	🟦 72.53	🟩 ~0.5
vMistral 7B Instruct		53.29	🟨 68.56	🟩 ~0.5
🟩 Green for the highest rank 🟦 Blue for second 🟨 Yellow for third				

06 Conclusion

Among the models tested, Claude 3.5 Sonnet exhibited the strongest overall performance. However, its superior capabilities came at the expense of slower generation speeds and higher costs. In contrast, Gemini 1.5 Flash emerged as a compelling option, balancing competitive performance with exceptional speed and cost-efficiency.

Open-source models, such as Mistral-7B-Instruct-V0.2 and Llama3-8B-Instruct, demonstrated their potential by delivering both faster generation times and lower costs compared to their proprietary counterparts. Notably, Mistral encountered difficulties with the MATH dataset, producing irrelevant outputs that included extraneous reasoning and question repetition instead of providing the required answers.

In terms of flexibility, open-source models can be a great option, as their performance can be enhanced through fine-tuning. While the initial costs – such as preparing a high-quality dataset for fine-tuning – may be higher, this approach becomes much more cost-effective and efficient in the long run in terms of both speed and cost.

Datasaur

datasaur.ai