A comparison of Large Language Models

5th June 2025

Contents

Comparison of Latest Large Language Models (LLMs): Price and Performance

Introduction

The rapid evolution of large language models (LLMs) has significantly impacted various fields, offering enhanced capabilities in natural language processing and understanding. This report aims to provide a comprehensive comparison of the latest LLMs, focusing on their price and performance metrics. By analyzing data from the past two months, we will evaluate models such as DeepSeek-R1, GPT-4.5, Claude-3.5-sonnet, Gemini 2.5 Pro Experimental, and Google Gemini-2.5-Pro. The objective is to identify which models offer the best value in terms of cost-effectiveness and performance efficiency, providing insights for stakeholders and researchers in the field.

Overview of Latest LLMs

The latest large language models (LLMs) under review include DeepSeek-R1, GPT-4.5, Claude-3.5-sonnet, Gemini 2.5 Pro Experimental, and Google Gemini-2.5-Pro. Each of these models represents a significant advancement in the field of natural language processing, offering unique features and capabilities that cater to different user needs and applications. DeepSeek-R1, released on January 20, 2025, is noted for its competitive performance and cost-effectiveness, aiming to deliver capabilities comparable to larger models at a reduced cost A Comparison of DeepSeek and Other LLMs. GPT-4.5, launched on February 27, 2025, focuses on energy efficiency with a projected 30% reduction in energy consumption Gemini 2.5 Pro vs GPT-4.5 - DocsBot AI. Claude-3.5-sonnet is recognized for its superior performance compared to DeepSeek, although specific release details are not provided A Comparison of DeepSeek and Other LLMs. Gemini 2.5 Pro Experimental and Google Gemini-2.5-Pro are also part of this comparison, with the former being highlighted for its high-quality output and a context window of 1 million tokens Gemini Pro - Google DeepMind. These models are evaluated based on their performance metrics, cost, and release dates to provide a comprehensive overview of the current landscape in LLM technology.

Comparison of Performance

Output Speed and Latency

The output speed and latency of large language models (LLMs) are critical factors in determining their efficiency and usability. Gemini 2.5 Pro Experimental demonstrates an impressive output speed of 161.4 tokens per second, which is indicative of its capability to handle complex tasks efficiently Gemini 2.5 Pro Experimental - Intelligence, Performance & Price .... In contrast, the time to first token (TTFT) for this model is 39.10 seconds, which may impact its responsiveness in real-time applications Gemini 2.5 Pro Experimental - Intelligence, Performance & Price .... These metrics highlight the trade-offs between speed and latency that users must consider when selecting an LLM for specific tasks. While Gemini 2.5 Pro Experimental excels in output speed, its latency could be a limiting factor in scenarios requiring immediate responses.

Qualitative Performance Insights

The qualitative performance of the latest large language models (LLMs) reveals distinct strengths and user experiences. Gemini 2.5 Pro is praised for its advanced reasoning and coding capabilities, making it suitable for complex tasks and leading in common benchmarks Gemini Pro - Google DeepMind. Users have noted its ability to process extensive datasets effectively due to its 1-million token context window, which enhances its performance in handling large-scale data Gemini Pro - Google DeepMind.

DeepSeek-R1 has been highlighted for its ability to solve complex coding problems that other models, such as ChatGPT, struggled with, indicating its strong problem-solving capabilities A Comparison of DeepSeek and Other LLMs. This model is particularly noted for its cost-effectiveness, providing competitive performance at a lower price point A Comparison of DeepSeek and Other LLMs.

GPT-4.5, while not as detailed in its performance metrics, is recognized for its energy efficiency, which is a significant consideration for users concerned with sustainability Gemini 2.5 Pro vs GPT-4.5 - DocsBot AI. This model's clear cost structure also aids users in budgeting for its use Gemini 2.5 Pro vs GPT-4.5 - DocsBot AI.

Overall, these models offer a range of capabilities that cater to different user needs, from high-performance reasoning and coding to cost-effective solutions and energy efficiency. The choice of model depends largely on the specific requirements of the task and the user's priorities in terms of cost, performance, and environmental impact.

Comparison of Pricing

The pricing structures of the latest large language models (LLMs) vary significantly, reflecting their diverse capabilities and target applications. GPT-4.5 is priced at $75.00 per million tokens for input and $150.00 per million tokens for output, indicating a premium for its advanced features and energy efficiency Gemini 2.5 Pro vs GPT-4.5 - DocsBot AI. In contrast, DeepSeek-R1 offers a more economical option with a cost of $0.80 per million tokens, making it a cost-effective choice for users DeepSeek R1 vs Gemini 2.5 Pro Experimental (Mar' 25).

The Gemini 2.5 Pro model does not have detailed pricing information available, but it is positioned as a high-quality model, suggesting a potential premium pricing strategy Gemini Pro - Google DeepMind. Claude 3.7 Sonnet, another model in the comparison, is priced at $2.00 per million tokens, indicating a balance between performance and cost Gemini 2.5 Pro Experimental - Intelligence, Performance & Price ….

Overall, the pricing of these models reflects their performance capabilities and the strategic positioning of their developers in the competitive LLM market. Users must consider both the cost per million tokens and the total projected costs when selecting a model that aligns with their specific application needs and budget constraints.

Model Name	Price (USD per 1M Tokens) - Input	Price (USD per 1M Tokens) - Output	Context Window	Release Date	Reference ID
Gemini 2.5 Pro	Unavailable	Unavailable	1M tokens	March 25, 2025	Gemini 2.5 Pro vs GPT-4.5 - DocsBot AI
GPT-4.5	$75.00	$150.00	128K tokens	February 27, 2025	Gemini 2.5 Pro vs GPT-4.5 - DocsBot AI
DeepSeek R1	$0.80	N/A	128K tokens	N/A	DeepSeek R1 vs Gemini 2.5 Pro Experimental (Mar' 25)
Claude 3.7 Sonnet	$2.00	N/A	200K tokens	N/A	Gemini 2.5 Pro Experimental - Intelligence, Performance & Price …

Pricing and Feature Comparison for LLMs

Conclusion

In conclusion, the comparison of the latest large language models (LLMs) highlights significant differences in their price and performance metrics, which are crucial for stakeholders and researchers in the field. DeepSeek-R1 emerges as a cost-effective option, offering competitive performance at a significantly lower cost compared to industry standards DeepSeek R1 vs Gemini 2.5 Pro Experimental (Mar' 25). Its impressive problem-solving capabilities make it suitable for real-time applications A Comparison of DeepSeek and Other LLMs. On the other hand, GPT-4.5, while also priced at $5.6 million, commands a higher price per million tokens due to its advanced features and energy efficiency Gemini 2.5 Pro vs

GPT-4.5 - DocsBot AI. Claude-3.7 Sonnet, although more expensive per token, offers superior performance in certain areas Gemini 2.5 Pro Experimental - Intelligence, Performance & Price …. The lack of detailed pricing for Gemini 2.5 Pro suggests a premium positioning in the market Gemini Pro - Google DeepMind. Overall, the choice of LLM should be guided by specific application needs, balancing cost with performance requirements.

References

1. Gemini Pro - Google DeepMind
2. Gemini 2.5 Pro Experimental - Intelligence, Performance & Price …
3. The Evolution of Large Language Models in 2024 and where we are …
4. LLM Pricing: Top 15+ Providers Compared in 2025
5. LLM Leaderboard - Compare GPT-4o, Llama 3, Mistral, Gemini …
6. A Comparison of DeepSeek and Other LLMs
7. Gemini 2.5 Pro vs GPT-4.5 - DocsBot AI
8. DeepSeek R1 vs Gemini 2.5 Pro Experimental (Mar’ 25)