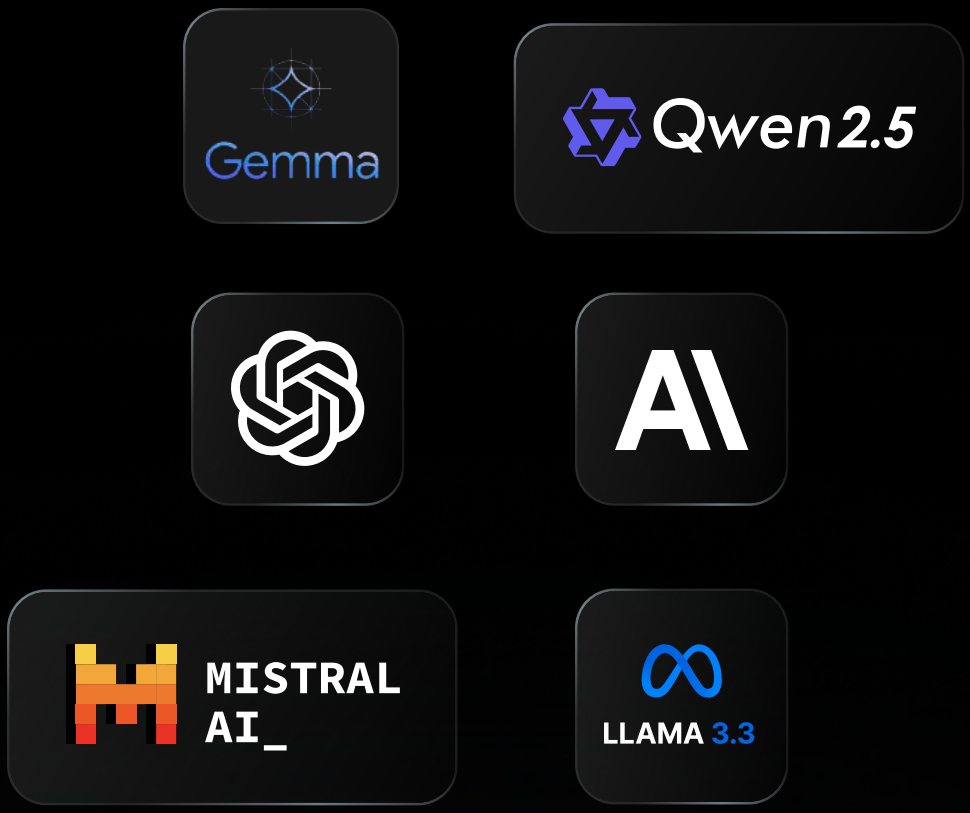
Datasaur



Choosing the Right LLM: An Exploration into How Different Models Stack Up in Performance

April 2024 | Whitepaper

01 Introduction

In the ever-evolving world of large language models (LLMs), there's always a buzz with new models popping up left and right, claiming to outshine the likes of OpenAI in different tasks. This flood of announcements often leaves data scientists scratching their heads, trying to make sense of it all. To cut through the confusion, we rolled up our sleeves and did a deep dive into each model's strengths and weaknesses. This whitepaper takes a closer look at the performance disparities among LLMs, uncovering what sets them apart and why.

02 Prominent Large Language Models in 2024

As we step into the year 2024, the artificial intelligence landscape is significantly shaped by the evolution and implementation of numerous notable Large Language Models (LLMs). To identify the top-performing LLMs, one can turn to the LMSys Chatbot Arena. This benchmark platform, designed for LLMs, hosts anonymous, randomized competitions in a crowdsourced format, providing a comprehensive and unbiased ranking of LLMs.

From this platform, we have identified six influential models that are leading the pack: GPT-4, Claude 3, Mistral, Llama-2, Gemma, and Qwen1.5-chat. In the following sections, we will delve into each of these models, discussing their unique strengths and contributions to the field.

GPT-4: Developed by OpenAI, GPT-4 is a large multimodal model designed to accept both text and image inputs and generate text outputs. It outperforms GPT-3.5-Turbo and other large language models (LLMs) in various benchmarks. It also excels in tasks that require visual inputs, such as generating captions, classifications, and analyses from images.

Claude 3: Claude 3 is a new family of large language models (LLMs) developed by Anthropic, an AI startup. Claude 3 was recently released with claims in the press release to have set new industry benchmarks across a wide range of cognitive tasks. The Claude 3 family includes three state-of-the-art models in ascending order of capability: Haiku, Sonnet, and Opus.

Mistral-7B: Introduced by Mistral.ai, the Mistral-7B is a pre-trained generative text model with 7 billion parameters and outperforms Llama 2 13B on various tasks. This model uses Grouped-Query Attention (GQA) and Sliding-Window Attention (SWA).

Llama-2: Developed by Meta AI, Llama 2 is an open-source LLM that is freely available for research and commercial purposes. This model is available in three sizes: 7B, 13B, and 70B. In benchmark comparisons, the largest Llama 2 has demonstrated competitive performance, outperforming other open-source LLMs like GPT-3.5 and PaLM on various tasks.

Gemma: Google Gemma is a new open-source AI model developed by Google. It was introduced to provide developers with advanced tools to create AI applications. Gemma AI is a lightweight, state-of-the-art open model with exceptional performance at its 2B and 7B sizes, but it faces challenges in complex reasoning and tracking objects, and its performance can vary by platform and implementation.

Qwen1.5-chat: Qwen-1.5 is a language model developed by QwenLM Team. It is a decoder-only transformer model with SwiGLU activation, RoPE, multi-head attention, and other features. Qwen-1.5 supports six model sizes: 0.5B, 1.8B, 4B, 7B, 14B, and 72B. It features significant improvements in chat model quality, strengthened multilingual capabilities, and system prompts that enable roleplay.

03 Large Language Model Evaluation

In this section, we will mainly examine the differences between utilizing OpenAI's proprietary LLM and open-source LLMs based on relevant criteria. OpenAI serves as our prime example of a proprietary LLM, given its exceptional performance and widespread popularity.

LLM Benchmark Dataset

The LLM benchmark dataset is a collection of datasets used to evaluate the performance of Large Language Models (LLMs). These datasets are designed to test various aspects of LLM performance, such as language understanding, generation, and manipulation. The benchmark datasets we used include:

MMLU (Measuring Massive Multitask Language Understanding): A dataset designed to test LLMs in various subjects on a variety of difficulty levels, ranging from elementary to professional level difficulty.

TruthfulQA: TruthfulQA is a benchmark dataset designed to measure the truthfulness of language models in generating answers to questions. The dataset comprises 817 questions that span 38 categories, including health, law, finance, and politics. The questions are crafted to test whether models can avoid generating false answers learned from imitating human texts.

MedQA: MedQA is a large-scale open domain question answering dataset from medical exams. It is collected from the professional medical board exams and covers three languages: English, simplified Chinese, and traditional Chinese. The dataset contains 12,723, 34,251, and 14,123 questions for the three languages, respectively. The dataset is used for multiple choice question answering.

LegalBench: LegalBench is an ongoing open science effort to collaboratively curate tasks for evaluating legal reasoning in English large language models (LLMs). The benchmark currently consists of 162 tasks gathered from 40 contributors. LegalBench offers a platform for researchers to evaluate the legal reasoning capabilities of LLMs and provides a means for the legal community to assess the performance of different LLMs for law-relevant tasks.

LegalSupport: This dataset assesses fine-grained reverse entailment. Each sample comprises a text passage presenting a legal claim and two case summaries, each describing a legal conclusion from different courts. The objective is to identify which case most strongly supports the legal claim in the passage. Annotations from a legal taxonomy inform the construction of this benchmark, distinguishing various levels of entailment. The task involves multiple-choice questions with two choices per question.

Experiment Setup

In our experiment, we utilize the Stanford Holistic Evaluation of Language Models (HELM) library, initiated by Lee and colleagues, to assess various Language Learning Models (LLMs). This library developed by the Stanford Center for Research on Foundation Models (CRFM), aims to enhance transparency in language models. It is regularly updated with new scenarios, metrics, and models, thanks to the collaborative efforts of the wider AI community.

HELM provides all the necessary components for running evaluations. Our modifications to the setup include adding models and setting the max-eval-instances to 100. Furthermore, we use the entire MMLU subset for evaluation, as opposed to the HELM benchmark's use of only 5 out of 57 subsets.

In this experiment, we assessed the performance of LLMs across various sizes to see whether there were notable improvements as the model size increased.

Experiment Results

After setting up the experiment with the aforementioned modifications, we initiated the experiment evaluations. **GPT-4-0613 emerged as the top performer**, boasting an average score of 0.787. It excelled in the MMLU, TruthfulQA, and MedQA categories, scoring 0.816, 0.86, and 0.846 respectively. In the LegalSupport category, it scored 0.6, which, while lower than its other scores, is still competitive when compared to other models. This suggests that GPT-4-0613 is a strong choice for a wide range of tasks, including those requiring legal support.

Model	Average	MMLU	TruthfulQA	MedQA	LegalBench	LegalSupport
Claude 3 Haiku	0.5958	0.726	0.65	0.75	0.533	0.32
Claude 3 Sonnet	0.6848	0.77	0.64	0.808	0.646	0.56
Claude 3 Opus	0.7328	0.766	0.78	0.769	0.709	**0.64**
Gemma-7B-IT	0.3862	0.282	0.24	0.346	0.423	**0.64**
GPT-3.5-Turbo (0125)	0.5524	0.687	0.59	0.635	0.45	0.4
GPT-4-0613	**0.787**	**0.816**	**0.86**	0.846	**0.813**	0.6
GPT-4 Turbo (0125 preview)	0.756	0.813	**0.86**	**0.865**	0.722	0.52
Llama-2-7B-Chat-HF	0.435	0.492	0.29	0.346	0.427	0.62
Llama-2-70B-Chat-HF	0.5606	0.634	0.5	0.558	0.551	0.56
Mistral-7B-Instruct-v0.1	0.4972	0.563	0.33	0.442	0.511	**0.64**
Mixtral-8x7B-Instruct-v0.1	0.6304	0.719	0.68	0.635	0.618	0.5
Qwen1.5-7B-Chat	0.376	0.608	0.17	0.423	0.199	0.48
Qwen1.5-72B-Chat	0.6956	0.773	0.76	0.769	0.656	0.52

Table 1: Evaluation results on several benchmark datasets from multiple instruction-following LLMs

Diving deeper into the performance of non-GPT models, we observed significant variation across different categories. For instance, the Gemma-7B-IT model, despite a lower average score of 0.3862, demonstrated remarkable performance in the LegalSupport category with a score of 0.64, suggesting its potential in legal support tasks. The Qwen1.5-72B-Chat model, with an impressive average score of 0.6956, excelled in the TruthfulQA and MedQA categories, making it a strong contender for tasks requiring truthful and medical-related responses. The Llama-2-7B-Chat-HF model, despite its lower average score of 0.435, showed promise in the LegalSupport category with a score of 0.62, indicating its potential for legal-related tasks. The Llama-2-70B-Chat-HF model, with an average score of 0.5606, showed a balanced performance across all categories, indicating its potential for tasks requiring multi-modal language understanding and its applicability in legal-related tasks.

The Mistral-7B-Instruct-v0.1 and Mixtral-8x7B-Instruct-v0.1 models, with average scores of 0.4972 and 0.6304 respectively, also performed admirably in the LegalSupport category, suggesting their potential applicability in legal-related tasks. Lastly, the Claude 3 Opus model, with a high average score of 0.7328, demonstrated consistent performance across all categories, indicating its adaptability to a wide range of tasks.

In conclusion, while GPT-4 may have the highest average score, other models also exhibit strong performance in specific categories and may be more suitable depending on the task requirements.

From Table 1 above, it can be observed that there is a general trend of **better performance with increasing model size**. For instance, the Claude 3 Opus model, which is larger than the Claude 3 Sonnet and Claude 3 Haiku models, shows better performance across all the tasks. Similarly, the GPT-4-0613 model, which is larger than the GPT-3.5-Turbo (0125), exhibits superior performance in all the tasks. This trend is also evident in the Llama-2-70B-Chat-HF model, which outperforms the smaller Llama-2-7B-Chat-HF model. The Qwen1.5-72B-Chat model also shows better performance than the smaller Qwen1.5-7B-Chat model. These observations suggest that larger models tend to have better performance, possibly due to their ability to capture more complex patterns and relationships in the data. However, it's important to note that this is a general trend and there may be exceptions based on specific tasks or datasets. This quality improvement also comes with tradeoffs in inference time and cost, which are outside the scope of this white paper.

While the general trend suggests that larger models tend to perform better, there are instances where smaller models outperform their larger counterparts. A notable example from the table is the Mixtral-8x7B-Instruct-v0.1 model, which performs better than the larger Llama-2-70B-Chat-HF model. This could be attributed to various factors such as the specific architecture of the model or the training data used.

Evaluation Technique

The evaluation results of language models are subject to several factors that can contribute to variations between attempts. Two primary factors influencing model output are (a) the specific prompt used to solicit a response and (b) the nature of the questions posed. For instance, differences may arise when requesting a binary label (0 or 1) compared to asking for labels expressed in words (e.g., spam or not spam). Similarly, obtaining a direct output from a model versus extracting an answer in a multiple-choice scenario can yield disparate results.

There are potential risks and impacts of inappropriately using evaluation benchmarks in the context of Large Language Models (LLMs). This issue is called benchmark leakage, where data related to evaluation sets is used for model training, leading to inflated and unreliable assessment of model performance.

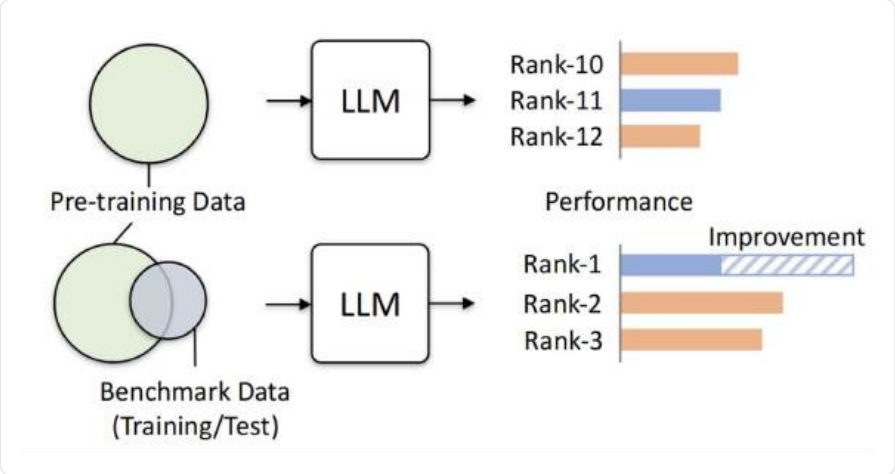


Fig. 1: Illustration of the potential risk of data leakage in paper written by Zhou, et al. Data leakage can boost benchmarking performance.

In the domain of evaluation, careful consideration must be given to how performance metrics are presented. These metrics can be sensitive to factors such as dataset splits and slight variations in inputs. While conducting multiple tests for each sample, employing diverse samples would be ideal. Consequently, it is paramount to approach reported numbers with a degree of caution and maintain a healthy level of skepticism regarding their interpretability and generalizability.

04 Selecting the Ideal Large Language Model (LLM)

Just as there is no universal code repository that fits all needs, there is no universally superior Large Language Model (LLM). Each use-case is distinct, and every organization operates within its unique set of parameters. The essence of a data scientist's role is to pinpoint the most appropriate model or architecture that is custom-fit to their specific circumstances. This can be achieved by:

- Starting with an examination of public evaluations, such as the leaderboards on Hugging Face (HF).
- Evaluating the model's performance across a range of benchmark datasets.
- Validating the model against a custom evaluation dataset that is in line with your use-case.
- It is recommended to initially start with smaller LLMs to save on both cost and inference time. If the results are not up to par, consider increasing the model size.
- Always remember to assess the model against other vital metrics such as cost and average inference time, particularly as you consider scaling the model to production-level use cases.

05 Conclusion

In conclusion, the field of Large Language Models (LLMs) is dynamic and ever-evolving, with a constant influx of new models, each boasting unique strengths and weaknesses. Our comprehensive exploration into the performance of various LLMs, including GPT-4, Claude 3, Mistral, Llama-2, Gemma, and Qwen1.5-chat, has revealed that while some models may excel in general tasks, others may shine in domain-specific tasks, underscoring the importance of task-specific model selection.

It's important to remember that there is no one-size-fits-all LLM. The ideal model is largely dependent on the specific use-case and the unique parameters within which an organization operates. Therefore, data scientists should initiate their selection process with public evaluations and validate the model against evaluation datasets that align with their use-case. While larger models generally tend to perform better, it's advisable to start with smaller LLMs and consider scaling up the model size as needed.

The world of LLMs is persistently advancing, necessitating regular re-evaluation of existing models to ensure they remain at the forefront of performance and relevance. A strategic investment in the ability to generate dedicated evaluation datasets, customized to each use case within your organization, can yield significant long-term benefits. This approach ensures that the chosen LLM is perfectly tailored to meet the unique needs and complexities of your specific applications, thereby maximizing its potential and effectiveness.

Datasaur, with its extensive selection of over 120 foundation models, offers a robust platform for data scientists to explore and select the ideal model for their specific use-case. Moreover, Datasaur provides comprehensive evaluation metrics, including quality, cost, and inference time, thereby facilitating a meticulous and reliable performance assessment. Additionally, Datasaur's advanced features for dataset generation provide users with the capability to create dedicated evaluation datasets, customized to use cases within their organization. These evaluation datasets can be run at a regular cadence, effectively serving as “unit tests” and protecting your model from regressing. Thus, Datasaur serves as a valuable tool in the dynamic and ever-evolving field of Large Language Models, aiding in the selection, customization, and evaluation of models to maximize their potential and effectiveness.

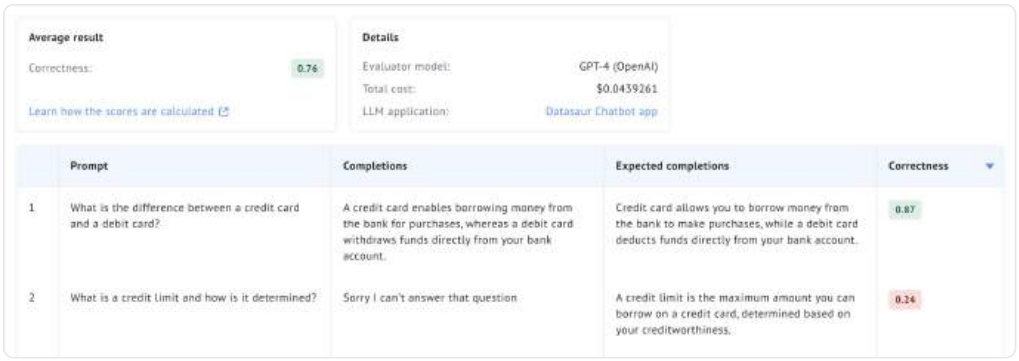


Fig 2: Evaluation feature in Datasaur

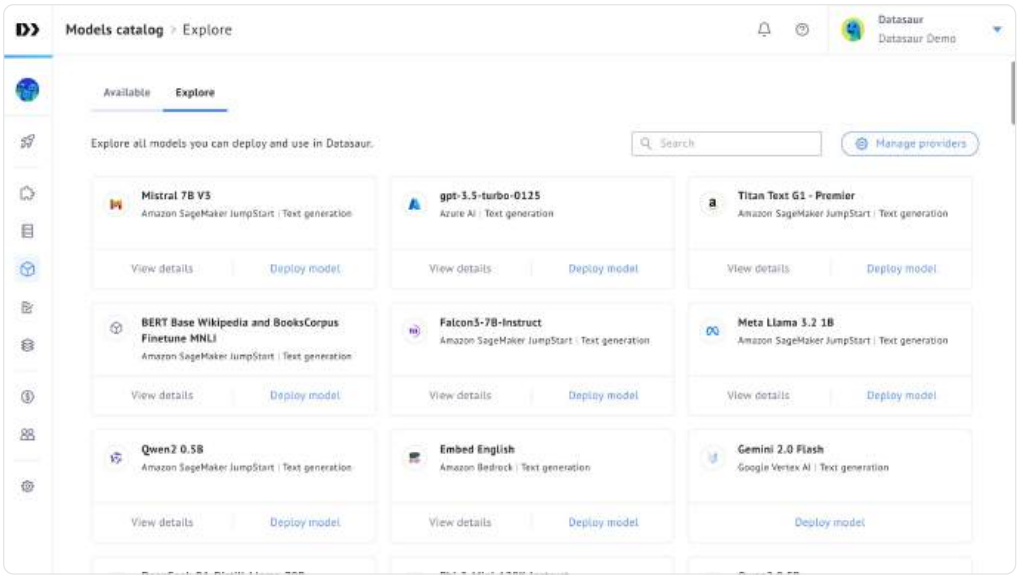


Fig 3: Datasaur provides more than 200 LLMs to choose

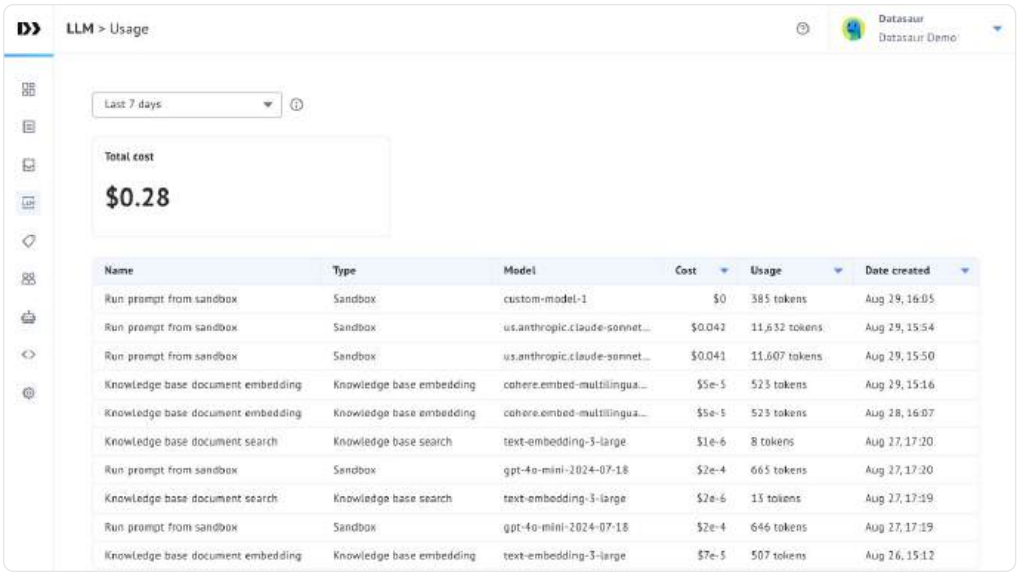


Fig 4: Cost tracking feature for each model in Datasaur

06 References

Lee, Tony, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, et al. "Holistic Evaluation of Text-To-Image Models." arXiv, November 7, 2023. https://doi.org/10.48550/arXiv.2311.04287.

Lee, Katherine, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. "Deduplicating Training Data Makes Language Models Better." arXiv, March 24, 2022. https://doi.org/10.48550/arXiv.2107.06499.

Zhou, Kun, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. "Don't Make Your LLM an Evaluation Benchmark Cheater." arXiv, November 3, 2023. https://doi.org/10.48550/arXiv.2311.01964.

About Datasaur

Datasaur is a private LLM provider and data labeling platform designed for companies to build their AI ecosystem with ease and efficiency. It assists organizations and universities in setting up custom LLMs and annotating data more efficiently and accurately through automation, quality control, and human-in-the-loop workflows. For more information, visit www.datasaur.ai.

Schedule a demo