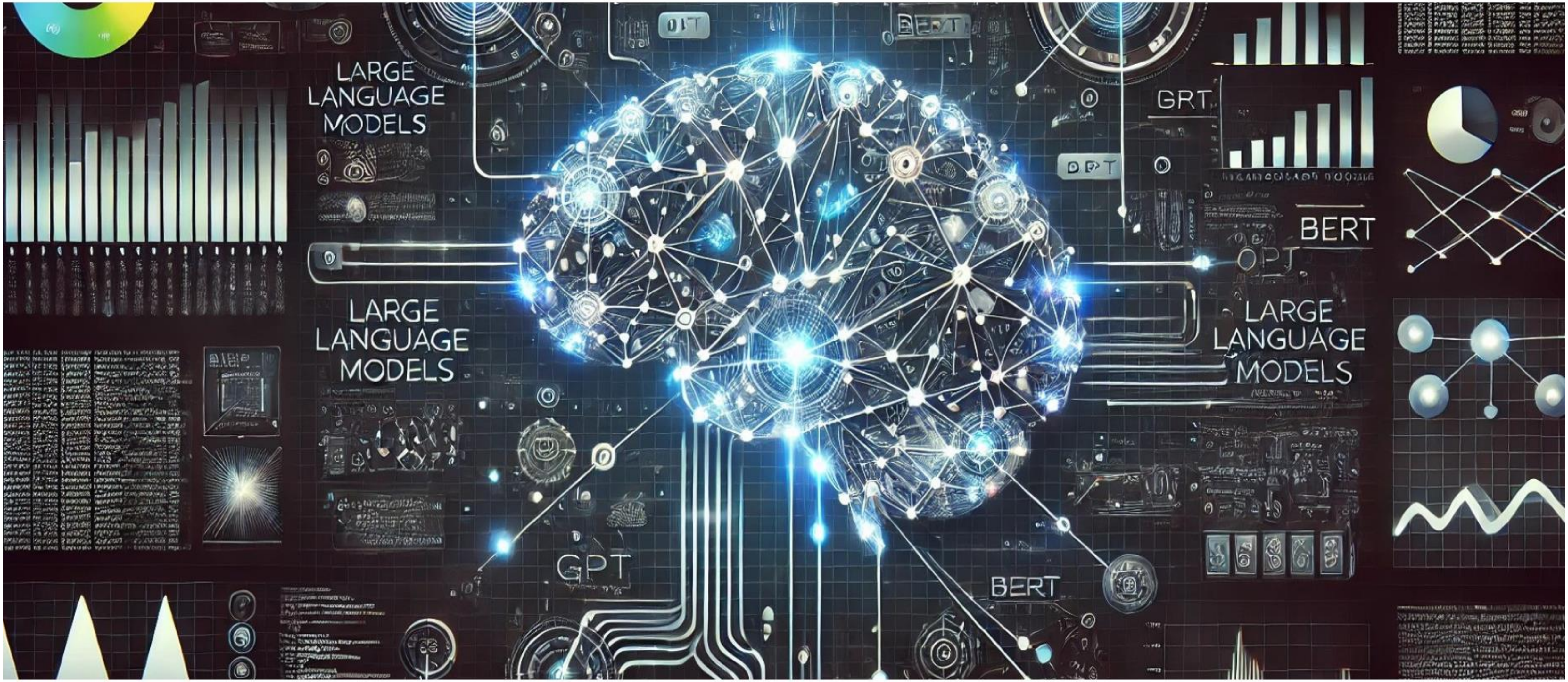
Benchmark Studio



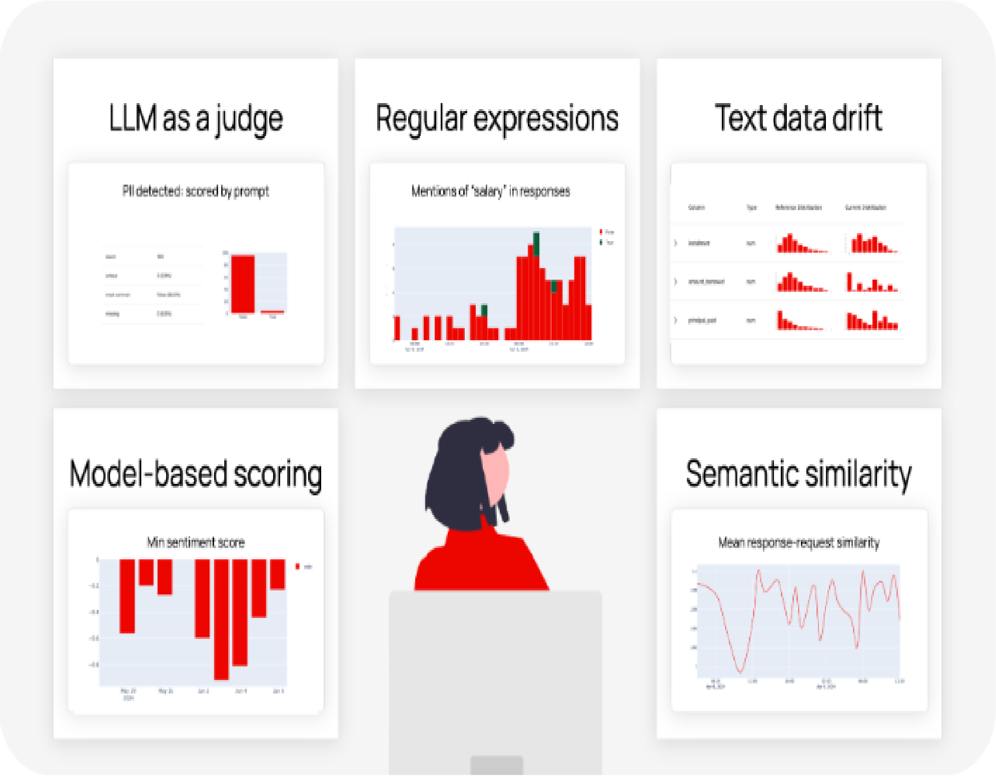
LLM Models Performance Comparison Analysis

Introduction

Nature of the Project

Currently, Large Language Models constitute the center of Generative Artificial Intelligence, which is a fast-growing area. These models understand text, voice, or image prompts and generate content with each prompt accordingly. The project aims to find the best LLM for a specific task's performance by analyzing the benchmark results for various LLMs. This will mainly focus on the factors and how each contributes to the performance of LLMs using exploratory data analysis. This would be done using publicly available data from various research papers, model repositories, and benchmark reports. Quite a few AI models have been developed in this era of AI, especially by giants like Google, OpenAI, Amazon, and Meta. Choosing the best among them has hence been an exhausting task. This analysis will minimize the time a user has to spend selecting a model for themselves; instead, it will get them the best pick that suits the work at hand.

It will also predict which model will last longer and give better responses, considering past evidence and other variables. It is expected that from this analysis, valuable insight will be gained into the efficiency and scalability of various LLM architectures, hence contributing to developing more efficient and resource-effective models. Comparisons between open-source and proprietary models will also be explored.



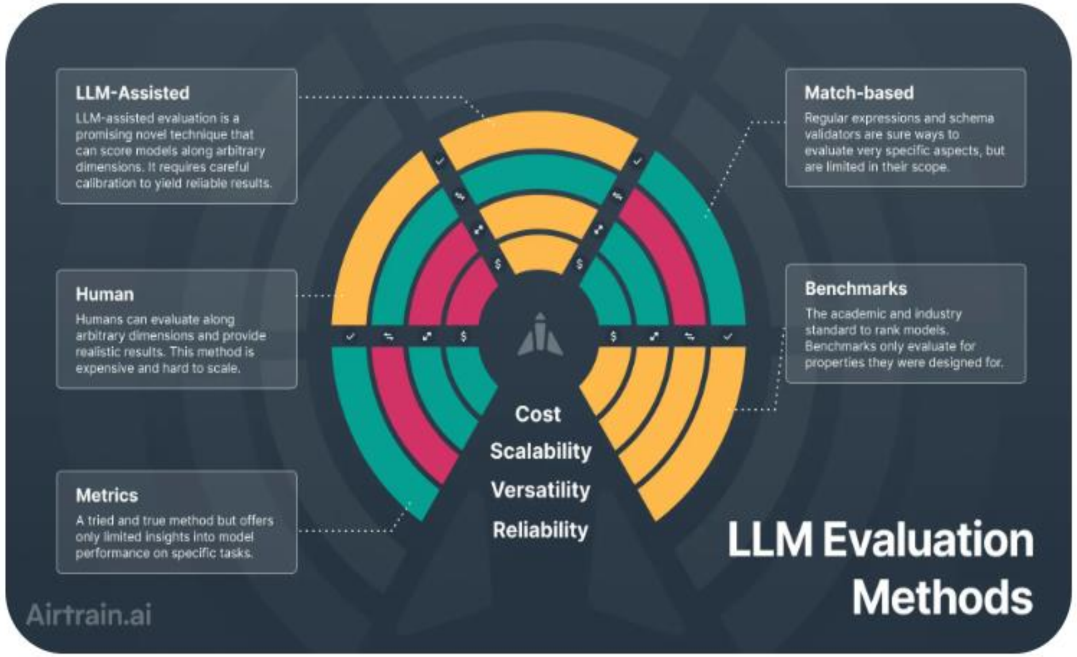
Why's it important?

Appropriate model selection is considered a challenge to tasks in rapid LLM development. Proper LLM model selection will reduce the time and cost of development associated with using a GPU. Comprehension of the performance trends in LLMs will help us understand where they perform well and poorly. A critical task will be finding hidden patterns in performance metrics data. Therefore, This optimal model may be pitted against various industry standards to find specific tasks for which it's optimized, narrowing down the most robust and efficient models. Since multiple architectures, such as Encoder-Decoder models, Transformers, or LSTMs, are being used in different LLMs, the question remains as to which one works best for a particular application. While the subtlety would be known to machine learning engineers and AI scientists, ordinary users of AI tools would know little about which model fits their application; hence, such a use case needed to be developed.

However, for an enterprise, any generative AI model provisioned for their daily actions should be to serve and maximize profitability. Thus, one needs to be more considerate of various factors and results of benchmarking when choosing the suitable AI model for use within a business instead of recognizing the brand. This helps identify significant pain points of LLMs for future consideration to develop the same.

Who's affected?

This analysis will enable data scientists and AI engineers to recognize their weaknesses and work on them for better outputs. It shall help businesses make informed decisions regarding the models that best suit their products. The same thing would be helpful at the academic level to provide insights from previous models and theories to the researchers and students. More accurate enhancements of the LLM models shall solve big and complicated problems, specifically in healthcare, where significant solutions can be provided. They will safely improve the interaction with customers in financial institutions, besides aiding in assessing and making decisions on risks. For the startups, they reduce the hustle of workflows by automating tasks, hence reducing workers. Developers can integrate large language models into AI applications, improving user experience and profitability, particularly within e-commerce and consultancy. Such analysis will also help the open-source community seek solutions to such models amidst various day-to-day problems. It will also enable them to inspire new ideas for startups.



LLM-Assisted
LLM-assisted evaluation is a promising novel technique that can score models along arbitrary dimensions. It requires careful calibration to yield reliable results.

Human
Humans can evaluate along arbitrary dimensions and provide realistic results. This method is expensive and hard to scale.

Metrics
A tried and true method but offers only limited insights into model performance on specific tasks.

Match-based
Regular expressions and schema validators are sure ways to evaluate very specific aspects, but are limited in their scope.

Benchmarks
The academic and industry standard to rank models. Benchmarks only evaluate for properties they were designed for.

What has been done so far and what gaps remain?

While a few tools or methodologies exist to compare the few popularly used LLMs, most need to be improved in predictability. Therefore, the project aims to identify the available LLMs and analyze the performance metrics across various benchmarks: complex datasets involving model architectures and training parameters that are often not available in a well-structured format. Most analyses so far need more comparisons regarding efficiency, accuracy, scalability, model size, and token size. Current models also do not address the language-dependent capabilities of LLMs. Few studies tried testing models varied for generative AI capabilities: adversarial robustness, coding, following instructions, and math and language skills. These involve, for example, manual evaluations by various prompts or subjective metrics that may introduce bias and limit generalizability. Instead, a more systematic and data-driven approach is required to evaluate LLMs, which will give better insights into performance and reduce the time spent executing the same prompt on various models. Addressing these limitations, this project will be able to provide critical insights into the choice of LLM and its performance.

Queries Related to Dataset

1. On what attributes does the prediction of LLM performance depend?
2. How does an LLM behave with respect to different languages utilized in a conversation?
3. Does the model perform radically different while generating code snippets than general text?
4. How often and for what reasons do models provide wrong answers?
5. How does the performance of an LLM vary with respect to the turn number within a prompt?
6. How are the prompt tokens related to the LLM performance?
7. In what way is the performance of LLMs different from deduplication tags?
8. Which models exhibit incremental improvement on all task categories?
9. Is any bias observed in model performance with respect to the language or category of a task?
10. Are there any models that are most likely to reject the tasks, and for what reason - due to improved overall performance or ethical reasons?

Our Team

Akshara Sri Lakshmipathy

Github LinkedIn

With a passion of uncovering hidden insights from data, I'm a dedicated my skills to drive impactful solutions. My academic background in AI and Machine Learning, combined with hands-on experience and project management, equips me to excel in this field. From building recommendation systems to predicting air pollution levels, my projects have solved real-world problems. I'm eager to contribute my skills to a dynamic and collaborative environment, where I can continue to learn, grow, and make a meaningful impact. My goal is to leverage data science to create innovative solutions that benefit society and drive business success.

Harish Nandhan Shanmugam

Github LinkedIn

I am a Data Science enthusiast with a strong background in Artificial Intelligence and Machine Learning. I enjoy analyzing data to uncover actionable insights and have experience using tools like Excel, Power BI, and Tableau for data visualization. My technical skills include managing SQL and NoSQL databases, cloud computing with AWS, and version control with Git. I am passionate about applying advanced data science techniques to solve real-world problems and drive organizational success.

Shivaraj Senthil Rajan

Github LinkedIn

I'm passionate about data science, where I explore the synergy between mathematics, statistics, and computer science. My focus is on artificial intelligence, machine learning, and data mining, aiming to leverage data's potential to drive innovation and scientific discovery. Beyond my technical expertise, I excel in team leadership, guiding projects and fostering collaboration to achieve collective goals. I'm dedicated to making a significant impact in the tech world, driven by a deep commitment to transforming data into actionable insights.