VISUALIZING PERFORMANCE PATTERNS OF OPEN-SOURCE LARGE LANGUAGE MODELS

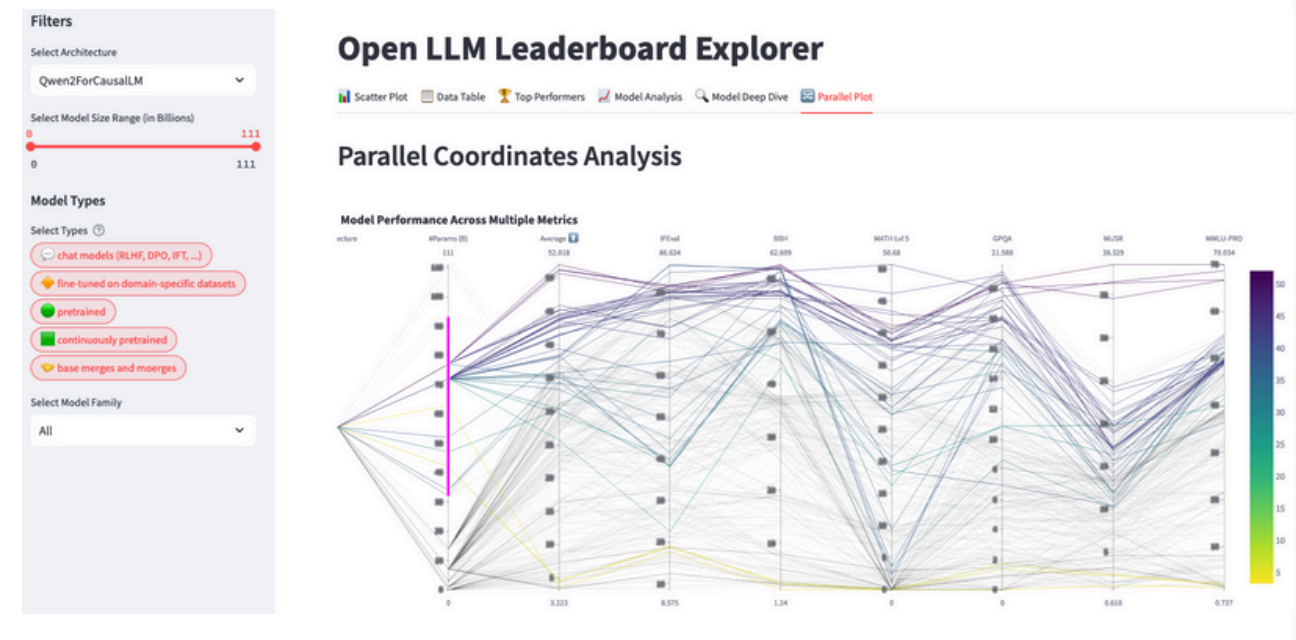
Authors: Lars Schmid (schmila7@students.zhaw.ch)
Katsiaryna Mlynchyk (mlynckat@students.zhaw.ch)
Supervisor: Prof. Dr. Susanne Bleisch (susanne.bleisch@fhnw.ch)
Date: 8 December 2024

ABSTRACT

We developed an interactive Streamlit app to explore the Open LLM Leaderboard dataset, enabling users to filter models, create visualizations, and analyze performance metrics. A strong positive correlation between model size and performance was revealed. The app guided the creation of infographics on selected models for specific GPU configurations and understanding the environmental trade-offs of scaling up.

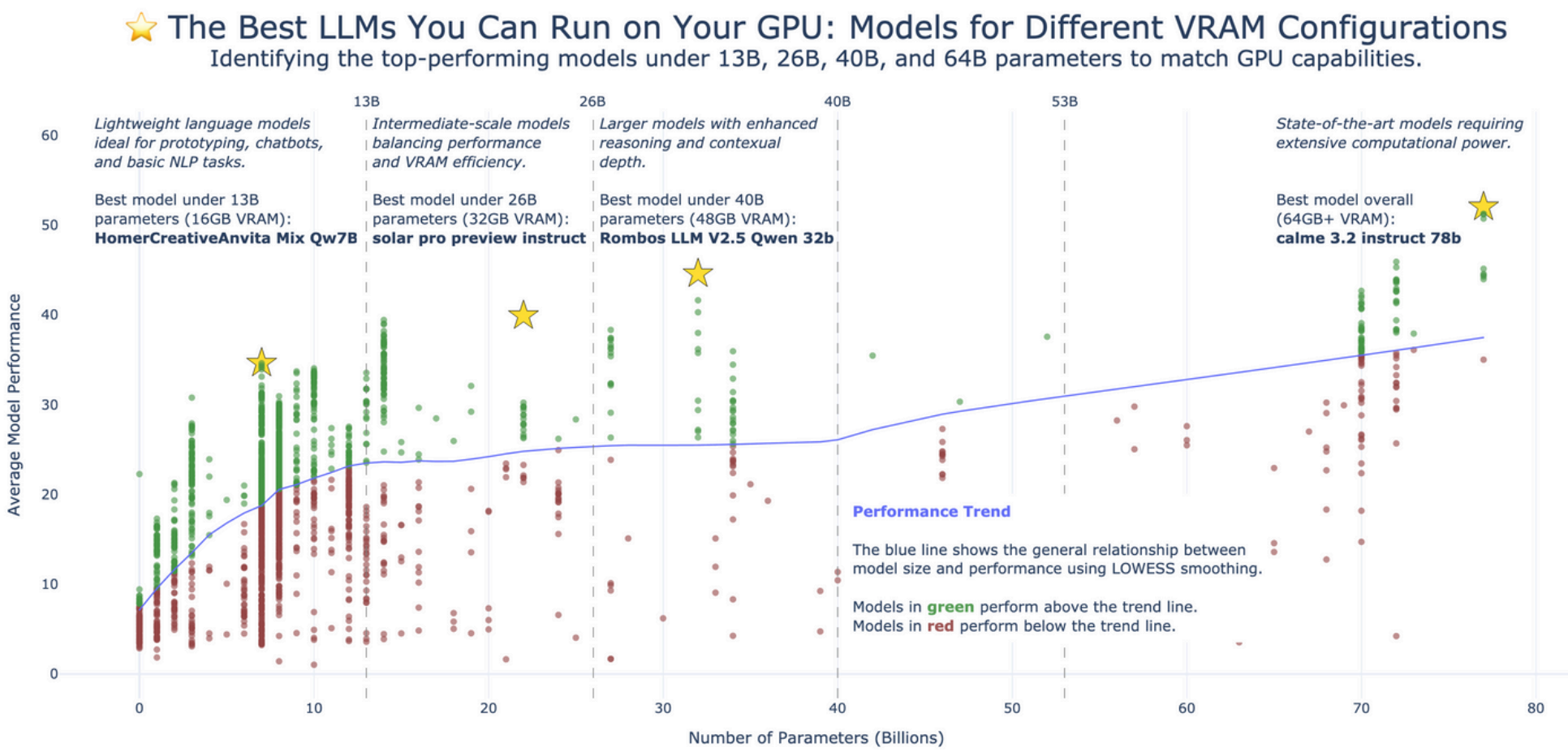
EXPLORATION & JOURNEY

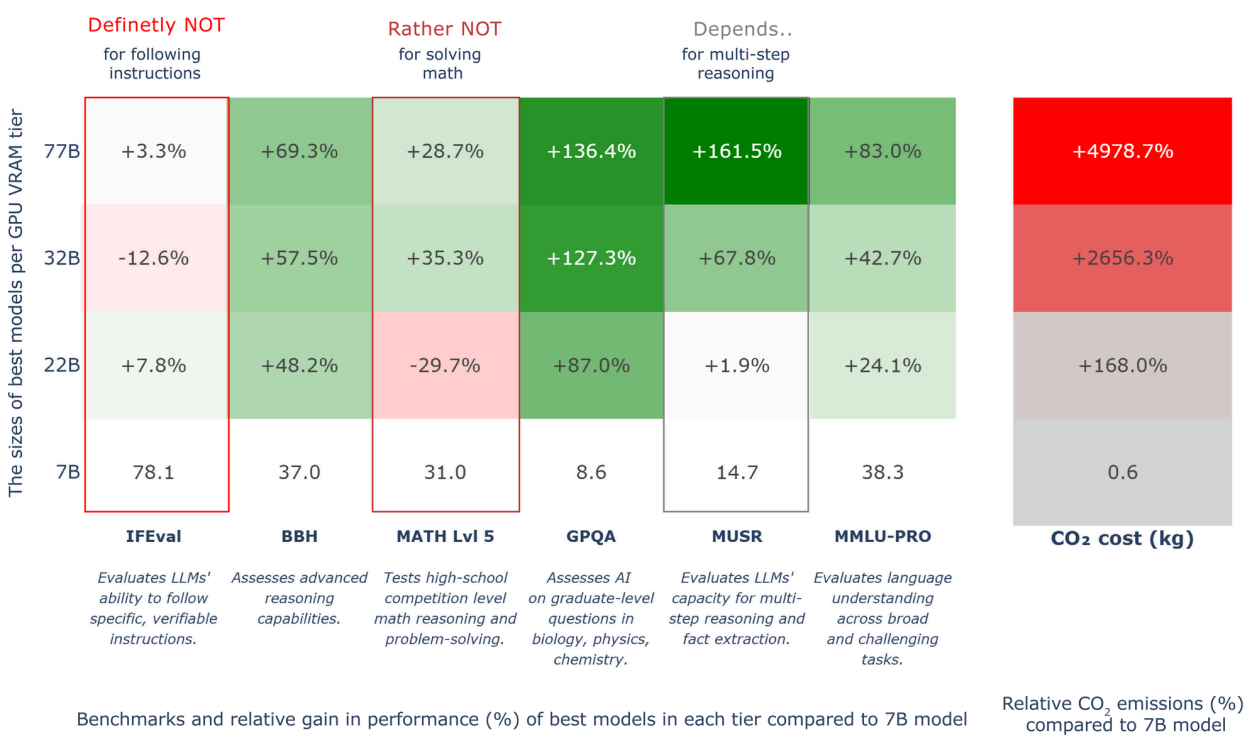
Using Python, Pandas, and Plotly, we developed an interactive Streamlit app to analyze the Open LLM Leaderboard dataset, enabling dynamic data exploration. The app includes several key features, such as filters to sort models by architecture, size, type, or family. It allowed us to create interactive visualizations, including scatter plots, radar charts, heatmaps, and parallel coordinate plots. Additionally, the app provided insights through statistical summaries, z-score normalization, and clustering techniques, offering a deeper understanding of the data. This tool played a central role in guiding our exploration and uncovering patterns while validating hypotheses.



The Open LLM Leaderboard dataset shows performance metrics for large language models. Data exploration revealed patterns in model size, performance, and efficiency. Using the Streamlit App for interactive analysis, two key findings emerged: 1) model parameters strongly correlate with performance across metrics, confirming larger models perform better; 2) performance gains plateau with size increases, showing diminishing returns. Further the deeper investigation of the best performing models and costs of running such models was conducted, resulting in two information graphics.

The first infographic provides a practical guide to model selection based on available GPU VRAM, highlighting the best-performing models for specific parameter thresholds.





The second infographic takes a more fine-grained look at the best models. How much performance in which tasks do you gain when you increase the size of the model? And how much more CO2 does it cause? Is it worth it? In most of the cases the answer is No.

LEARNING

Reflecting on this task, we learned the importance of dedicating time to thorough data exploration and the surprising complexity of creating impactful infographics, which demand far more attention to detail than typical plots. The development of an interactive app has made it much easier for us to identify patterns and test hypotheses and demonstrates the value of interactive tools for data analysis.

SOURCES

Hugging Face. Open LLM Leaderboard. Retrieved from https://huggingface.co/datasets/open-llm-leaderboard/contents
Substratus AI. Calculating GPU memory for serving LLMs. Retrieved from https://www.substratus.ai/blog/calculating-gpu-memory-for-llm
Mlynchyk, K., & Schmid, L. Open LLM Leaderboard Explorer [Streamlit app]. Retrieved from https://github.com/mlynckat/InfVis_HS24