IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

Ranking and Comparison Tool for Large Language Models

Dr. G. Lakshmi Narayana[1], Sahithi[2], M. Ramu[3], M. Sandhya[4], P. Harshith[5]
Associate Professor, Department of CSE[1]
UG Students, Department of CSE[2-5]
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh.
lakshminarayana.gumma@gmail.com, 22bq1a05i3@vvit.net, 22bq1a05d9@vvit.net
22bq1a05c8@vvit.net, 22bq1a05g6@vvit.net

Abstract: *Large Language Models (LLMs) have rapidly evolved and are now widely used across applications such as chatbots, content generation, coding assistance, and data analysis. With the increasing number of LLMs available, selecting the most suitable model for a specific task has become a significant challenge. Existing evaluation benchmarks often focus on isolated metrics and do not adequately capture real-world user preferences or comparative performance. This project presents Ranking and Comparison Tool for Large Language Models, a system designed to provide an intuitive, transparent, and dynamic method for evaluating and ranking LLMs. The proposed system employs an Elo rating based ranking mechanism, widely used in competitive systems, to rank models through pairwise comparisons. Models are evaluated based on multiple qualitative and quantitative dimensions such as accuracy, helpfulness, creativity, and response quality. An interactive user interface allows users to compare models visually, filter rankings based on task requirements, and explore performance trends over time. Experimental evaluation demonstrates that the Elo-based approach produces stable and interpretable rankings that adapt effectively to new model updates and user feedback. The system bridges the gap between traditional benchmark-based evaluation and human-centric assessment. Future enhancements include automated benchmark integration, expert-weighted evaluations, and real-time ranking updates.*

Keywords: LLM ranking, Elo rating, model comparison, interactive evaluation, benchmarking

I. INTRODUCTION

Large Language Models (LLMs) are increasingly being adopted by students for learning, problem solving, explanation generation, and conceptual understanding across a wide range of academic subjects. Despite their widespread availability, LLMs differ significantly in terms of reasoning capability, response clarity, cost efficiency, multimodal support, and tool integration. As a result, selecting an appropriate model for educational use has become a non-trivial task for both students and institutions. To address this challenge, the OpenTests project was initiated as a structured evaluation platform aimed at identifying which LLMs perform best in educational workflows. The initial objective of the platform was to benchmark models using objective metrics related to correctness and representational quality. However, as the system evolved, it became evident that output variability across repeated executions posed a challenge to static evaluation methods. This observation led to the development of a arena-based evaluation mechanism that complements traditional benchmarking. OpenTests now unifies controlled evaluation, real-world student preference feedback, and transparent model metadata into a single, comprehensive evaluation ecosystem designed specifically for educational environments.

DOI: 10.48175/IJARSCT-32407

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

II. RELATED WORKS

The evaluation of Large Language Models (LLMs) has traditionally been carried out using benchmark-driven methodologies, where models are assessed against fixed datasets using metrics such as accuracy, reasoning performance, and language understanding. While these benchmarks offer standardized and reproducible evaluation, recent studies have pointed out their limitations in capturing real-world usage patterns and qualitative aspects of model responses. In particular, the non deterministic nature of generative models often leads to output variability, making one-time benchmark evaluations insufficient to represent long-term model behavior. To overcome the limitations of static evaluation, researchers have explored human-in-the-loop and pairwise comparison approaches. In these methods, evaluators compare responses from multiple models and select the preferred output, enabling better assessment of subjective qualities such as clarity, coherence, and usefulness. Pairwise evaluation has been shown to produce more stable rankings than absolute scoring systems, especially when combined with aggregation techniques such as the Elo rating system. Elo-based ranking has been successfully adapted from competitive domains to AI model evaluation, allowing rankings to evolve dynamically as new comparison data becomes available. Public leader boards and crowdsourced evaluation platforms have further extended model comparison by incorporating large scale user feedback. However, many such systems lack transparency in evaluation criteria, model metadata, and operational constraints such as cost and tool support. Additionally, reliance solely on user voting can introduce bias and overlook objective correctness. These gaps in existing literature motivate the need for evaluation frameworks that balance automated benchmarking with human preference based ranking while transparency and adaptability.

III. METHODOLOGY

System Overview

The proposed **Ranking and Comparison Tool for Large Language Models** is a modular evaluation platform designed to assess LLMs using both automated testing and user preference analysis. The system enables users to compare responses generated by different models through an A/B testing interface while keeping model identities hidden to ensure unbiased evaluation. Internally, the platform supports tool-assisted LLM inference and automated evaluation using an LLM-as-a-judge approach. Model performance is aggregated using an **Elo-based ranking mechanism**, which dynamically updates rankings as new comparison data becomes available. The resulting rankings and model metadata are presented through an interactive dashboard to assist users in selecting appropriate models for specific tasks.

The **OpenTests platform** integrates multiple Large Language Models to enable comparative evaluation of responses generated by different systems. The platform currently supports several models from different providers, including **GPT-OSS-120B, GPT-OSS-20B, Kimi-K2 Instruct, Qwen-3-32B, Groq Compound, Groq Compound Mini, Llama-4 Maverick, and Llama-4 Scout**. These models represent diverse architectures and capabilities, enabling meaningful comparison within the arena-based evaluation environment. To maintain consistent interaction with the supported models, responses are obtained through a unified inference interface. The system retrieves model outputs through an API-based inference service implemented using the **Groq API**.

Evaluation Strategy

We wanted to answer a simple but important question: **Which AI model should students and educators trust with their learning?**

Our Two Golden Tests

In education, there are really two things that matter most when evaluating an AI model:

Finding the Right Information: Does the model give correct, accurate answers? Can it solve problems correctly?

Representing Information the Right Way: Even if the answer is correct, can the model explain it in a way that's easy to understand? Does it use visuals, diagrams, and clear explanations?

These became our two golden test datasets:

Raw Intelligence: Tests whether models can find the right answers to educational questions.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

Visualization: Tests whether models can represent information in helpful, visual ways.

The Problem with Benchmarks

However, challenges arise when relying solely on benchmark-based evaluation.We ran these tests and got answers. But we quickly realized something important: Benchmarks alone do not tell the whole story.

The results kept changing. Different questions, different contexts, different ways of measuring—everything affected the outcomes. What seemed like a "better" model one day might score differently the next. We weren't alone in this experience. The entire AI community faces this challenge.

Arena-Based Evaluation Approach

To overcome the limitations of static benchmark-based evaluation, the OpenTests platform introduces an arena-based evaluation approach. In this framework, users submit a query and are presented with responses generated by two different Large Language Models while the model identities remain hidden. Users compare the responses side-by-side and select the response they consider more useful or accurate. Each comparison contributes to a pairwise evaluation that updates model rankings using an Elo-based scoring mechanism. This approach enables the evaluation system to incorporate real user preferences rather than relying solely on predefined benchmarks.

The arena-based evaluation provides several advantages:

Democratic evaluation: Users participate directly in comparing model responses, allowing the ranking system to reflect collective user preferences.

Transparent decision process: The comparison mechanism and ranking updates are visible and understandable, making the evaluation process more transparent.

Inclusive feedback: Different users may prioritize different qualities such as accuracy, clarity, or depth of explanation, allowing the system to capture diverse perspectives.

Arena-Based Public Evaluation To capture real-world variability and evolving user preferences, the proposed system incorporates an arena-based public evaluation mechanism. In this mode, users submit educational queries and are presented with responses generated by two anonymized Large Language Models. Users select the response they find more useful or accurate, thereby providing preference-based feedback.

These user selections are continuously aggregated and used to update model rankings dynamically. Unlike static benchmarks, this approach captures model consistency over time, student preference trends, and practical usefulness in real educational scenarios. While arena-based rankings evolve based on ongoing user interactions, private benchmark scores are retained as baseline references to provide contextual grounding for the rankings.

3.3 Ranking Mechanism

The proposed system employs an **Elo-based ranking mechanism** to rank Large Language Models based on pairwise comparison outcomes. Initially, each model is assigned a baseline rating. During arena-based evaluation, two anonymized models are selected and their responses to the same user query are compared. When a user selects a preferred response, the selected model is considered the winner of the comparison, and the opposing model is treated as the loser. The Elo framework updates model ratings by comparing the expected outcome, derived from existing ratings, with the actual outcome of the comparison.

The rating adjustment is controlled using a configurable update factor to balance ranking stability and adaptability. To avoid volatility due to insufficient data, ranking updates are applied only after a minimum number of comparisons have been recorded for each model. Rankings are periodically recalculated to account for newly added models and updated versions. The resulting Elo scores are used to generate live leader boards and visual summaries, allowing users to observe performance trends over time while ensuring fair and reliable model comparison.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

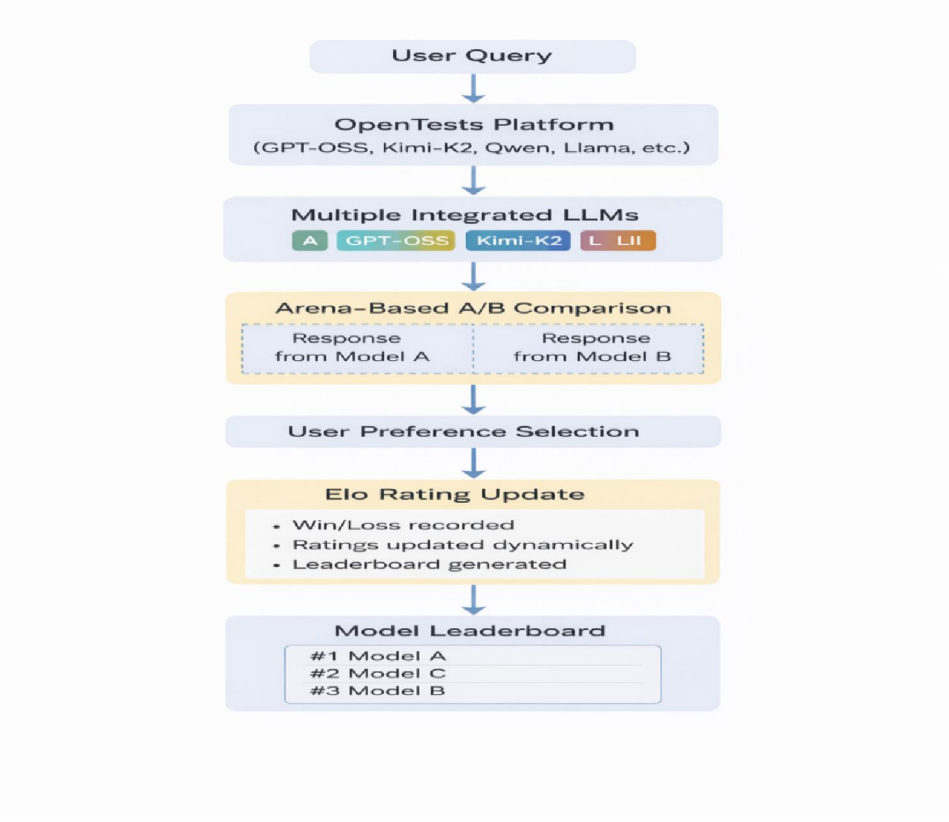


Fig. 1. Arena-based A/B comparison process used in the OpenTests platform.

IV. RESULTS AND DISCUSSION

The public arena evaluation module was deployed and tested with multiple Large Language Models across a series of user-driven pairwise comparisons. Users were presented with anonymous responses from two models and asked to select the better answer based on educational usefulness, clarity, and correctness. The system records wins, losses, and updates model rankings using an ELO-based scoring mechanism.

Figure 2 illustrates the distribution of wins and losses across evaluated models. The visualization demonstrates clear performance differentiation, with certain models consistently achieving higher win counts, while others exhibit a higher loss frequency. This indicates that the arena mechanism is capable of capturing comparative model quality even with limited samples.

DOI: 10.48175/IJARSCT-32407

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

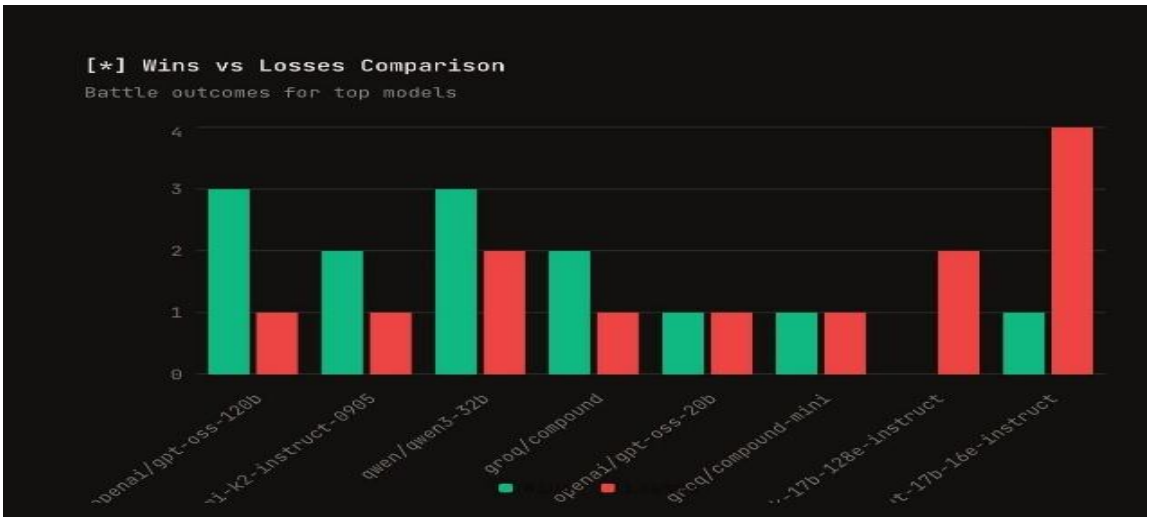
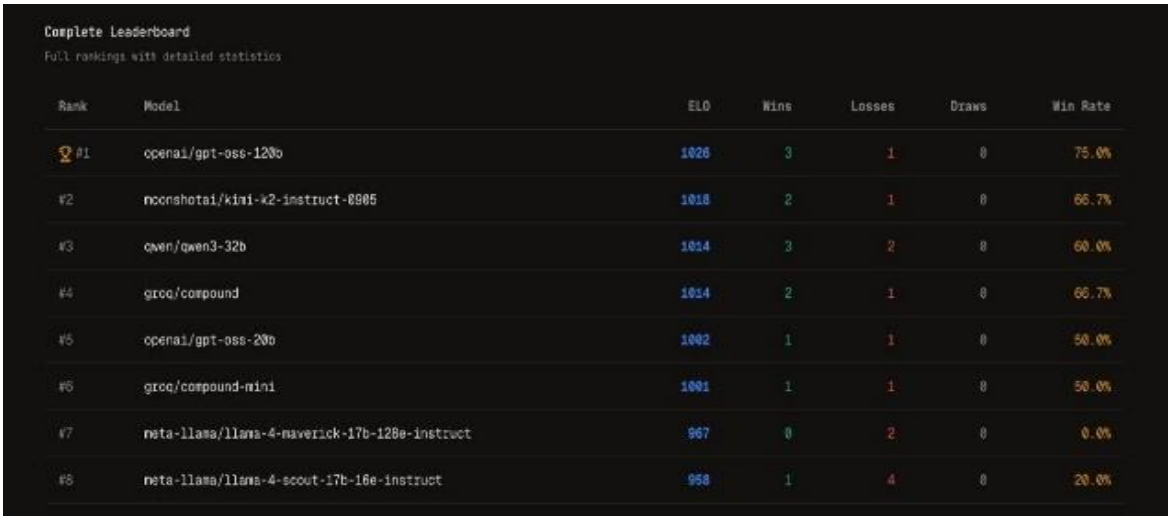


Fig. 2. Distribution of wins and losses across evaluated models.

Figure 3 presents the complete leaderboard generated by the arena system, including ELO scores, win rates, and match statistics. The top-ranked model achieved a win rate of approximately 75%, while mid-tier models maintained win rates between 50–67%. Lower-ranked models demonstrated significantly reduced win rates, validating the sensitivity of the ranking mechanism to user preferences and response quality.



Complete Leaderboard

Full rankings with detailed statistics

Rank	Model	ELO	Wins	Losses	Draws	Win Rate
#1	openai/gpt-oss-120b	1026	3	1	0	75.0%
#2	moonshotai/kimi-k2-instruct-0905	1018	2	1	0	66.7%
#3	qwen/qwen3-32b	1014	3	2	0	60.0%
#4	groq/compound	1014	2	1	0	66.7%
#5	openai/gpt-oss-20b	1002	1	1	0	50.0%
#6	groq/compound-mini	1001	1	1	0	50.0%
#7	meta-llama/llama-4-maverick-17b-128e-instruct	967	0	2	0	0.0%
#8	meta-llama/llama-4-scout-17b-16e-instruct	958	1	4	0	20.0%

Fig. 3. Arena-based leaderboard showing Elo scores and win rates.

Figure 4 shows the win-rate trend across ranked models. The downward progression reflects consistent ordering between visual analytics and leaderboard statistics, confirming metric stability and correctness of aggregation logic.

DOI: 10.48175/IJARSCT-32407

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

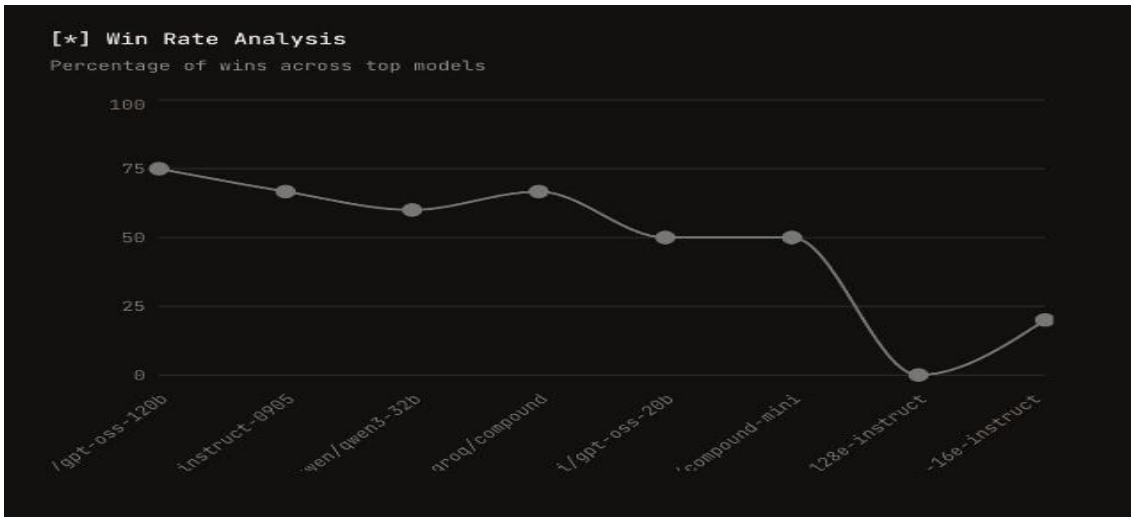


Fig. 4. Win-rate trend across ranked language models.

These results demonstrate that the arena-based evaluation successfully captures real-world user preferences, enables continuous ranking updates, and provides transparent comparative insights into model performance for educational usage.

V. CONCLUSION

OpenTests demonstrates a practical and scalable approach to assessing Large Language Models in educational environments. By combining automated benchmarking, arena-based human preference analysis, and transparent model metadata, the platform provides a comprehensive framework for model comparison and selection. The hybrid architecture effectively accommodates the inherent non-determinism of modern LLMs while maintaining objective quality assessment. As the adoption of LLMs continues to expand in education, OpenTests offers a reliable foundation for guiding users toward effective and trustworthy AI systems.

Furthermore, the proposed framework illustrates how integrating automated benchmarks with human preference–driven arena comparisons can provide a more balanced and reliable assessment of model performance. This approach enables continuous model comparison while adapting to evolving user needs and model updates. The system can also serve as a foundation for future research on transparent and user-centered AI model comparison platforms.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to their respected guide Dr. G. Lakshmi Narayana for the continuous support, valuable suggestions, and insightful guidance throughout the course of this work. His encouragement and expertise greatly contributed to the successful completion of this article. We are also thankful to the Project Coordinator, Dr. N. Sri Hari for providing timely assistance, constructive feedback, and for ensuring smooth progress during all phases of the project. Our heartfelt thanks go to the Head of the Department, Dr. V. Ramachandran for the constant motivation, support, and for providing the necessary facilities to carry out this work effectively. We extend our deep appreciation to the Principal, Dr. Y. Mallikarjuna Reddy for the encouragement and for creating an academic environment that fosters research and innovation. Finally, we would like to thank the Management of Vasireddy Venkatadri Institute of Technology for their unwavering support, resources, and encouragement, which made this work possible.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

REFERENCES

[1] L. Chen, Z. Qin, Y. Guo, J. Rohde, and Y. Zhang, "Benchmarking Large Language Models on Homework Assessment in Circuit Analysis," arXiv preprint arXiv:2506.06390, Jun. 2025.
[2] L. Wang, D. Yi, D. Jose, J. Passarelli, J. Gao, J. Leventis, and K. Li, "Enterprise Large Language Model Evaluation Benchmark," arXiv preprint arXiv:2506.20274, Jun. 2025.
[3] J. S. Jauhiainen and A. Garagorry Guerra, "Evaluating Students' Open-ended Written Responses with LLMs: Using the RAG Framework for GPT-3.5, GPT-4, Claude-3, and Mistral-Large," arXiv preprint arXiv:2405.05444, May 2024.
[4] S. Al Faraby, A. Romadhony, and A. Adiwijaya, "Analysis of LLMs for Educational Question Classification and Generation," Computers & Education: Artificial Intelligence, vol. 7, Dec. 2024, 100298.
[5] K. Busch and H. Leopold, "Towards a Benchmark for Large Language Models for Business Process Management Tasks," arXiv preprint arXiv:2410.03255, Oct. 2024.
[6] K. Chernyshev, V. Polshkov, E. Artemova, A. Myasnikov, V. Stepanov, A. Miasnikov, and S. Tilga, "U-MATH: A University-Level Benchmark for Evaluating Mathematical Skills in LLMs," arXiv preprint arXiv:2412.03205, Dec. 2024.
[7] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Zhang, and P. Liang, "Holistic Evaluation of Language Models," arXiv preprint arXiv:2211.09110, Nov. 2022.
[8] W. Zheng, L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, and I. Stoica, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," arXiv preprint arXiv:2306.05685, Jun. 2023.
[9] D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, and J. Steinhardt, "Measuring Massive Multitask Language Understanding," arXiv preprint arXiv:2009.03300, Sep. 2020.
[10] J. Srivastava, A. Rastogi, A. Rao, A. Reddy, and A. Mishra, "Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models," arXiv preprint arXiv:2206.04615, Jun. 2022.