Emerging Trends in LLM Benchmarking

Akshar Prabhu Desai
akshar.iitb@gmail.com

Ritu Prajapati
rituprajapati53@gmail.com

Tejasvi Ravi
ravitejasvi@gmail.com

Mohammad Luqman
luqmankgp@gmail.com

Pranjul Yadav
impranjul@gmail.com

2024 IEEE International Conference on Big Data (BigData) | 979-8-3503-6248-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/BigData62323.2024.10825845

***Abstract*—Traditionally, machine learning models that focused on specialized tasks facilitated straightforward evaluation. However, the evolution of Large Language Models, has increased the complexity w.r.t. performance measurement. Evaluation and benchmarking of large language model is a significant challenge due to their versatility and improved capability to perform a wide range of tasks. This manuscript examines existing literature for various benchmarks and identifies a comprehensive overview of the emerging trends in benchmarking methodology.**

***Index Terms*—Large Language Models, Machine Learning, Benchmarking**

I. Introduction

Benchmarking is the process of evaluating the performance of Large language models (LLM) against standard set of tasks. Benchmarking enables researchers to guide the development of the LLMs and evaluate its effectiveness against other similar competing models while giving insights into the strengths and weaknesses of the LLMs being evaluated.

While there is past literature [1] that covers different benchmarks and evaluation methodologies for LLM; LLMs have been improving rapidly in their capabilities and this necessitates a need to improve the benchmarks as well. Creating LLM benchmarks that are sufficiently hard to beat as well as comprehensive enough to enable model comparison has become a challenge in itself. In the subsequent sections we cover the emerging trends in development of LLM benchmarks and what we can expect in near future.

In Section II, we cover the emerging trends of comprehensive benchmarks which consider multi-dimensional assessments and real world tasks. In Section III, we present how LLM are compared for specific capabilities such as reasoning, tool usage or code analysis. Lastly, in section IV we present human centric benchmarking trends which focus on explainability, reasoning and evaluation with implicit feedback.

II. Benchmarking against comprehensive evaluation criteria

Traditionally, machine learning models were built to solve specific problem and hence evaluation was easier against specific tasks. LLMs, on the other hand, showed promise across multiple different tasks. Liang et al. [2] through their HELM benchmark showed that LLMs are good at wide variety of tasks across multiple domains and a holistic evaluation is necessary that goes beyond accuracy metrics. Such comprehensive evaluation can be classified under following subcategories.

A. Multi-dimensional assessment

Most recent benchmarks focused on LLMs have moved from a single score assessment to multi-dimensional assessment which measures performance of LLMs across multiple dimensions. White et al. [3] for example presents LiveBench, a benchmark that measures performances of LLMs across dimensions such as Coding, language, reasoning, instruction following, math and data analysis.

HELM [2] is another multi dimensional benchmark that creates a taxonomy of scenarios and metrics to provide a holistic evaluation framework. For example, for language specific tasks HELM has created 16 scenarios such as summarization, question answering, sentiment analysis etc. For metrics HELM focuses not just on correctness but also has measures of fairness, robustness and toxicity.

B. Real-world tasks

As LLMs are being increasingly applied to real world tasks, developing such real world task benchmarks is a new emerging trend. Reed et al. [4] shows GATO agent's adaptability to perform a variety of tasks and the need of multi-modal benchmarks to assess the agent's capabilities in real-world settings. Srivastava et al. [5] proposed a benchmark that consists of tasks covering diverse domains such as linguistics, mathematics, common sense reasoning, and social bias detection to evaluate LLM for real world tasks.

OSWorld [6], a benchmark to measure performance of LLMs and VLMs against real world computer usage tasks, reveals significant limitations of state-of-the-art LLM/VLM-based agents. While humans achieve a success rate of over 72%, the best performing model only reaches 12% [6].

Research suggests that LLMs are still very much behind humans when it comes to tasks in real-world settings hence benchmarks are crucial to improve the existing LLMs.

III. Benchmarking against specific capabilities

With evolving language models and its expanding real world application, one emerging trends in benchmarking is to measure performance of models for a specific capability. While traditional benchmarking can evaluate the model for accuracy, it may fail to address the strength or weaknesses of the model within a specific context.

Guo et al. [7] brings attention to the need of a comprehensive LLM evaluation with knowledge and reasoning as

979-8-3503-6248-0/24/$31.00 ©2024 IEEE
Authorized licensed use limited to: GOOGLE. Downloaded on February 03,2025 at 19:27:18 UTC from IEEE Xplore. Restrictions apply.

fundamental to the capabilities of LLM. LogiQA [8] is one such benchmark that goes beyond factual question answering and measures the performance of LLM based on its capability to answers through reasoning and deduction over an expert written dataset of question. ReClor [9] is another benchmark that explains how Machine reading comprehensive models can exploit the bias in the dataset to achieve high accuracy score without understanding the context. To challenge the model for its reasoning capabilities, ReClor evaluated the model against datasets with and without biases and exposed how pre-trained models performed poorly with random guessing on dataset without a bias. Highlighting, the need of more logical reasoning targeted benchmarks.

Code generation which is seen as a special ability of LLMs, has seen progress from simple benchmarks such as MBPP (Mostly Basic Python Problems) [10] which targets Python for its simplicity to benchmarks such as HumanEval [11] which can evaluate a model's capability to solve higher complexity problems.

As LLMs are deployed in specialized domains like medicine, science and legal etc, specialized benchmarks are even more necessary to make sure the model is reliable and fine-tuned for higher accuracy. MedQA [12] is a benchmark that evaluate the model against real world medical problems. LEGALBENCH citeguha2023legalbench benchmark evaluates a model for its legal reasoning via a set of tasks categorized among various legal reasoning.

Research around Specific capability benchmarks underlines the need of advanced benchmarks to evaluate these models for higher accuracy and guide development of next set of models.

IV. Human centric benchmarking

Language tasks are inherently subjective in nature due to context dependency, ambiguity, and reliance on human interpretation. While Automated metrics can capture the technical aspects or correctness, they often fail to capture the subjective qualities. Human centric benchmarking focuses on this aspect of LLM tasks and involves humans in evaluating a model's output, crafting datasets and formulating prompts. Following is a broad classification of ways human centric benchmarking can bridge the gap between traditional automatic metrics and the nuances of language.

A. Data Creation and Curation

Human involvement in designing, building or improving datasets ensures diversity and a broader range of language use. HumanEval [11] utilizes a dataset of hand-written coding problems designed to benchmark the code-writing capabilities of LLMs against a diversified and realistic problem scenarios. TruthfulQA [13] uses a set of human crafted questions that test truthfulness of an LLM model.

B. Task and prompt Development

LLM benchmarking with an emphasis on task and prompt development by humans reflects the understanding that humans are best equipped to create tasks even with the nuances of language understanding. Such tasks and prompts can also address any potential harm associated with LLM. Big-Bench [5] introduces a collaborative benchmark with tasks contributed by numerous authors. TruthfulQA [13] introduces a set of questions crafted by humans to evaluate if a model can avoid generating false answers learned from imitating human texts.

Evaluation tasks and prompt crafted by humans ensures that the evaluation is comprehensive, and the LLM performance is aligned and relevant to human use of a language.

C. Evaluation and Feedback

Humans can directly evaluate LLM outputs for qualities that automatic metrics often miss, such as coherence, readability in summaries, or potential harm in generated text. This evaluation can directly be used as a standard of LLM performance, to compare the evaluation with automatic metrics, or as a feedback to train the LLM model for fine tuning.

Stiennon et al. [14] showcases the role of human evaluator in ranking outputs which are provided as direct input for fine-tuning the models. The result shows that training with human feedback significantly outperforms strong baselines on English summarization. Gehman et al. [15] discusses a benchmark where humans evaluate the outputs for toxicity to ensure that models do not generate harmful or offensive content. Direct evaluation by humans can also include reviewing or rating the automated metrics based on their perceptions of output quality. This can help ensure that automated metrics accurately reflect human judgment and remain useful for benchmarking.

Human involvement in feedback and evaluation helps fine tune the model towards outputs that align with human preferences while also addressing limitations of relying solely on objective metrics.

D. Shaping Benchmarking Practices

Direct Human involvement during the benchmarking process can help make sure the process is robust and reliable given the fast moving evolution of LLMs. Stiennon et al. [14] showcases how human involvement can ensure reliable benchmarking results by establishing quality control measures. Human involvement can help identify the most relevant and useful metrics to assess the performance of LLM models and developing new metrics that are currently hard to assess with existing automated metrics.

BIG-bench [5] exemplifies the role of human collaboration in shaping benchmarking practices. Humans contribute various tasks and perspectives to this benchmark which allows a more broader and detailed evaluation of the models.

Even though helpful in numerous ways, human involvement creates challenges for scaling and accounting for the bias among the humans. We expect the future research to focus on developing scalable human centric benchmarks.

V. Conclusion

Benchmarking guides the development of LLMs and allows for comparison with other models. The emerging trends in benchmarking suggests three broad areas of interest for LLM

Authorized licensed use limited to: GOOGLE. Downloaded on February 03,2025 at 19:27:18 UTC from IEEE Xplore. Restrictions apply.

benchmarks. Firstly, comprehensive benchmarks that measure LLM performance across multiple dimensions with complex tasks. Secondly, specialize benchmarks that measure LLMs for their performance against highly specialized tasks. Lastly, as LLMs improve and match human performance in many areas, human centric evaluation criteria that uses real humans to evaluate LLMs is also an area of active research and interest.

REFERENCES

[1] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. 15(3), March 2024. ISSN 2157-6904. doi: 10.1145/3641289. URL https://doi.org/10.1145/3641289.

[2] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2023. URL https://arxiv.org/abs/2211.09110.

[3] Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-free llm benchmark, 2024. URL https://arxiv.org/abs/2406.19314.

[4] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A generalist agent, 2022. URL https://arxiv.org/abs/2205.06175.

[5] Aarohi Srivastava et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023. URL https://arxiv.org/abs/2206.04615.

[6] Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments, 2024. URL https://arxiv.org/abs/2404.07972.

[7] Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. Evaluating large language models: A comprehensive survey, 2023. URL https://arxiv.org/abs/2310.19736.

[8] Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning, 2020. URL https://arxiv.org/abs/2007.08124.

[9] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning, 2020. URL https://arxiv.org/abs/2002.04326.

[10] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models, 2021. URL https://arxiv.org/abs/2108.07732.

[11] Daniel Li and Lincoln Murr. Humaneval on latest gpt models – 2024, 2024. URL https://arxiv.org/abs/2402.14852.

[12] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *CoRR*, abs/2009.13081, 2020. URL https://arxiv.org/abs/2009.13081.

[13] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *CoRR*, abs/2109.07958, 2021. URL https://arxiv.org/abs/2109.07958.

[14] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc., 2020.

[15] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.301.

Authorized licensed use limited to: GOOGLE. Downloaded on February 03,2025 at 19:27:18 UTC from IEEE Xplore. Restrictions apply.