***Abstract*—**Gemma models has shown impressive capabilities on many benchmarks. However, the authors did not test other open source models against Gemma because their setting does not guarantee for fairness among all of them. Moreover, there are more specialized benchmarks on which the performance of Gemma is yet to be determined. In our proposal for GSoC 2025, we suggest coding a tool for evaluating a set of Text only and Multimodal Large Language Models.

1 INTRODUCTION

During the course of this project, I aim at developping a framework to evaluate Multimodal Large Language Models of the following features:

- Support for diverse set of benchmark datasets,
- Build an interface for Huggingface Models to support many models including Gemma, LLaMA, Mistral and other models released in the future.
- Easy interface to run on benchmark datasets, and
- A leaderboard including all results for models tested with our framework.

2 BENCHMARKS

There has been a lot of benchmarks introduced to test MLLMs. So, I select a sample of them to be representative to many categories from development perspective. For other benchmark testsets, they are expected to be added at a later stage to test the modularity of the framework.

2.1 Commonly Used Benchmarks

This is a list of commonly used benchmarks in reporting LLMs/MLLMs performance:

1. MMLU-PRO Wang et al., 2024 and/or MMLU-ProX Xuan et al., 2025
2. TruthfulQALin, Hilton, and Evans, 2021
3. HellaSwag Zellers et al., 2019 and/or HellaSwag-Pro Li et al., 2025
4. Big-Bench Lite authors, 2023
5. IFEval Zhou et al., 2023 and/or IFEval-Extended Kovalevskyi, 2024

2.2 Coding Tasks

- CodeXGLUE Lu et al., 2021

2.3 Chart Related Tasks

2.3.1 *Chart-to-Table*

- Testsets: ChartQA Masry et al., 2022, PlotQA Methani et al., 2020, ICPR22
- Metrics: Relative Number Set Similarity and Relative Mapping Similarity Liu et al., 2022

2.3.2 *Chart Question Answering*

- Testsets: ChartQA Masry et al., 2022, PlotQA Methani et al., 2020, ICPR22
- Metrics: Accuracy, Precision, Recall and F1

2.3.3 *Chart Summarization*

- Testsets: Chart-to-Text Kantharaj et al., 2022 and ChartSumm Rahman et al., 2023
- Metrics: BLEU Post, 2018, CIDEr Vedantam, Lawrence Zitnick, and Parikh, 2015, ROUGE Lin, 2004 and BLEURT Sellam, Das, and Parikh, 2020.

2.4 LLMs as Agents

- Software Engineering Agents: SWE-Bench Yang et al., 2024
- Machine Learning Researchers: MLAgentBench Huang et al., 2023, MLGym Nathani et al., 2025

2.5 Task Selection

The following is the list of selected benchmark to implement: TBD

We have two main componenets: MLLM evaluation and Leaderboard.

To illustrate the role for each componenet, assume the following scenario:

1. The researchers want to evaluate their model during Instruction Fine-tuning process or a model that was generated due to a previous tuning step. So, they use MLLM Evaluation tool via python function calling or command line interface, respectively.
2. If they provide a link to the Leaderboard, MLLM Evaluation tool communicates with the Leaderboard via http/https to send the results.
3. The researchers could track the tables in the Leaderboard and see their cells getting filled with evaluation results.

3 TECHNICAL IMPLEMENTATION PLAN

In the following subsections, I explain in details the steps to make the evaluation framework. In table **??**, you can find the list of tasks with their expected deadlines.

Task	Deadline
Models Unified Interface	
Command Line Interface	
Implement Leaderboard Interface	
Add support for other models and datasets	

Table 1—Tasks Timetable.

3.1 MLLM Evaluation Core

First, I will write down the main skeleton of the tool. This includes:

1. Implement the main function for Gemma 3 4B on single dataset,
2. taking arguments via both command line and function calling,
3. evaluating the model on one benchmark, and
4. reporting the results in form of printed text (output of CLI).

3.2 Add support for some models and datasets

Update the code to work with other models like LLaMA and Mistral, and the other datasets **??**.

3.3 Leaderboard

I plan to make the leaderboard a web application that shows the results of the evaluation for all models. Other evaluation scripts are expected to contact this app with the final scores via http/https requests. For frameworks, I plan to use FastAPI, Angular and Nginx.

4 DELIVERABLES

By the end of this project, I expect to deliver:

1. Evaluation Git Repo: repository containing the code for MLLM evaluation,
2. Leaderboard Repo: repository containing the code for Leaderboard web app, and

3. System Paper: explaining the overall system and its design choices.

5 ANTICIPATED IMPACT

By the end of this project, we end up with an MLLM evaluation tools that make it easier to evaluate MLLMs, compare among them and understand their capabilities in fine-grained details across various general and domain specific tasks.

6 REFERENCES

1. Lin, Chin-Yew (2004). "Rouge: A package for automatic evaluation of summaries". In: *Text summarization branches out*, pp. 74–81.
2. Vedantam, Ramakrishna, Lawrence Zitnick, C, and Parikh, Devi (2015). "Cider: Consensus-based image description evaluation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575.
3. Post, Matt (2018). "A call for clarity in reporting BLEU scores". In: *arXiv preprint arXiv:1804.08771*.
4. Zellers, Rowan, Holtzman, Ari, Bisk, Yonatan, Farhadi, Ali, and Choi, Yejin (2019). "Hellaswag: Can a machine really finish your sentence?" In: *arXiv preprint arXiv:1905.07830*.
5. Methani, Nitesh, Ganguly, Pritha, Khapra, Mitesh M, and Kumar, Pratyush (2020). "Plotqa: Reasoning over scientific plots". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1527–1536.
6. Sellam, Thibault, Das, Dipanjan, and Parikh, Ankur P (2020). "BLEURT: Learning robust metrics for text generation". In: *arXiv preprint arXiv:2004.04696*.
7. Lin, Stephanie, Hilton, Jacob, and Evans, Owain (2021). "Truthfulqa: Measuring how models mimic human falsehoods". In: *arXiv preprint arXiv:2109.07958*.
8. Lu, Shuai, Guo, Daya, Ren, Shuo, Huang, Junjie, Svyatkovskiy, Alexey, Blanco, Ambrosio, Clement, Colin, Drain, Dawn, Jiang, Daxin, Tang, Duyu, et al. (2021). "Codexglue: A machine learning benchmark dataset for code understanding and generation". In: *arXiv preprint arXiv:2102.04664*.
9. Kantharaj, Shankar, Leong, Rixie Tiffany Ko, Lin, Xiang, Masry, Ahmed, Thakkar, Megh, Hoque, Enamul, and Joty, Shafiq (2022). "Chart-to-text: A large-scale benchmark for chart summarization". In: *arXiv preprint arXiv:2203.06486*.
10. Liu, Fangyu, Eisenschlos, Julian Martin, Piccinno, Francesco, Krichene, Syrine, Pang, Chenxi, Lee, Kenton, Joshi, Mandar, Chen, Wenhu, Collier, Nigel, and Altun, Yasemin (2022). "Deplot: One-shot visual language reasoning by plot-to-table translation". In: *arXiv preprint arXiv:2212.10505*.

11. Masry, Ahmed, Long, Do Xuan, Tan, Jia Qing, Joty, Shafiq, and Hoque, Enamul (2022). "Chartqa: A benchmark for question answering about charts with visual and logical reasoning". In: *arXiv preprint arXiv:2203.10244*.
12. authors, BIG-bench (2023). "Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models". In: *Transactions on Machine Learning Research*. ISSN: 2835-8856. URL: https://openreview.net/forum?id=uyTL5Bvosj.
13. Huang, Qian, Vora, Jian, Liang, Percy, and Leskovec, Jure (2023). "Mlagentbench: Evaluating language agents on machine learning experimentation". In: *arXiv preprint arXiv:2310.03302*.
14. Rahman, Raian, Hasan, Rizvi, Al Farhad, Abdullah, Laskar, Md Tahmid Rahman, Ashmafee, Md Hamjajul, and Kamal, Abu Raihan Mostofa (2023). "ChartSumm: A Comprehensive Benchmark for Automatic Chart Summarization of Long and Short Summaries." In: *Canadian AI*.
15. Zhou, Jeffrey, Lu, Tianjian, Mishra, Swaroop, Brahma, Siddhartha, Basu, Sujoy, Luan, Yi, Zhou, Denny, and Hou, Le (2023). "Instruction-following evaluation for large language models". In: *arXiv preprint arXiv:2311.07911*.
16. Kovalevskyi, Bohdan (2024). "IFEval-Extended: Enhancing Instruction-Following Evaluation in Large Language Models through Dynamic Prompt Generation". In: *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023* 5.1, pp. 513–524.
17. Wang, Yubo, Ma, Xueguang, Zhang, Ge, Ni, Yuansheng, Chandra, Abhranil, Guo, Shiguang, Ren, Weiming, Arulraj, Aaran, He, Xuan, Jiang, Ziyan, et al. (2024). "Mmlu-pro: A more robust and challenging multi-task language understanding benchmark". In: *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
18. Yang, John, Jimenez, Carlos, Wettig, Alexander, Lieret, Kilian, Yao, Shunyu, Narasimhan, Karthik, and Press, Ofir (2024). "Swe-agent: Agent-computer interfaces enable automated software engineering". In: *Advances in Neural Information Processing Systems* 37, pp. 50528–50652.
19. Li, Xiaoyuan, Li, Moxin, Men, Rui, Zhang, Yichang, Bao, Keqin, Wang, Wenjie, Feng, Fuli, Liu, Dayiheng, and Lin, Junyang (2025). "HellaSwag-Pro: A Large-Scale Bilingual Benchmark for Evaluating the Robustness of LLMs in Commonsense Reasoning". In: *arXiv preprint arXiv:2502.11393*.
20. Nathani, Deepak, Madaan, Lovish, Roberts, Nicholas, Bashlykov, Nikolay, Menon, Ajay, Moens, Vincent, Budhiraja, Amar, Magka, Despoina, Vorotilov,

Vladislav, Chaurasia, Gaurav, et al. (2025). "MLGym: A New Framework and Benchmark for Advancing AI Research Agents". In: *arXiv preprint arXiv:2502.14499*.

21. Xuan, Weihao, Yang, Rui, Qi, Heli, Zeng, Qingcheng, Xiao, Yunze, Xing, Yun, Wang, Junjue, Li, Huitao, Li, Xin, Yu, Kunyu, et al. (2025). "MMLU-ProX: A Multilingual Benchmark for Advanced Large Language Model Evaluation". In: *arXiv preprint arXiv:2503.10497*.