ETH zürich

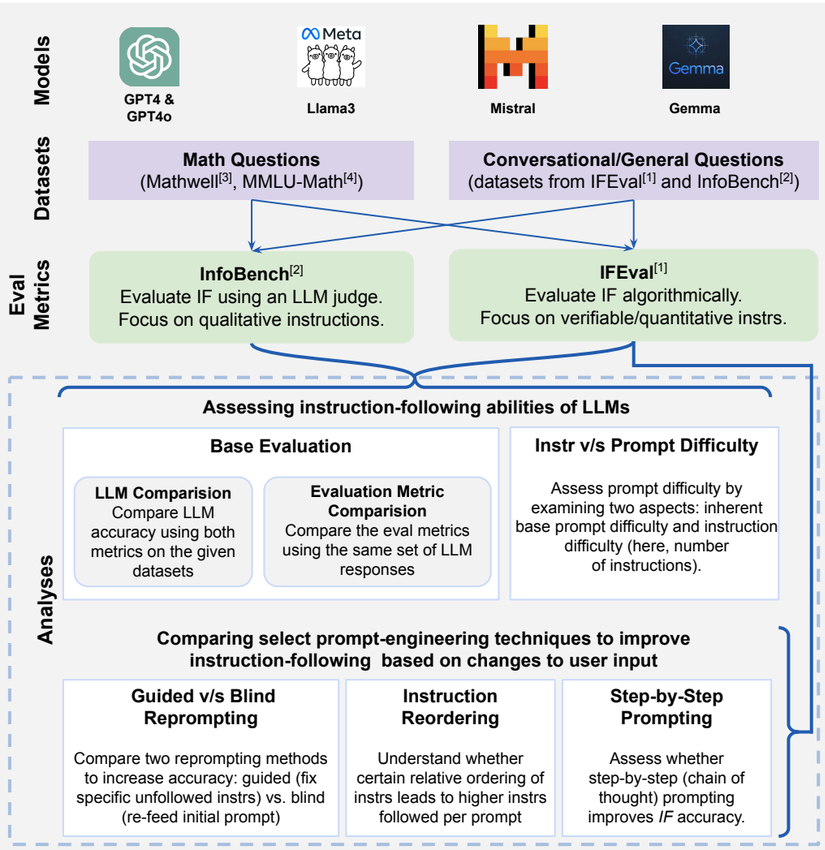
# Assessing Instruction Following Capabilities of LLMs

Baidyanath Kundu, Garima Singh, Piyushi Goyal
D-INFK, ETH Zurich, Switzerland

## 1 Introduction

- Large Language Models (LLMs) excel in problem-solving and Instruction-Following *(**IF**)*, making them valuable for many applications.
- As such, several research works have been dedicated to better understand the abilities of LLMs, especially in the context of *IF*.
- This project builds on existing works and performs a comprehensive study on the *IF* capabilities of popular LLMs through 5 different experimental analyses.

## 2 Experiments




## 3 Results

### 3.1 Base Evaluation

| | IFEval *IFEval* | InfoB *IFEval* | IFEval *InfoB* | InfoB *InfoB* |
|---|---|---|---|---|
| GPT4 | 67 | 64 | 61 | 74 |
| GPT4o | 66 | 51 | 63 | 80 |
| Llama 3 | 52 | 57 | 56 | 69 |
| Mistral | 53 | 43 | 54 | 46 |
| Gemma | 63 | 45 | 54 | 52 |

Fully closed-source LLMs take the lead. The GPT4 family greatly outperforms its open counterparts.

Within the open-source models, Llama3 performs the best, followed by Gemma and Mistral.

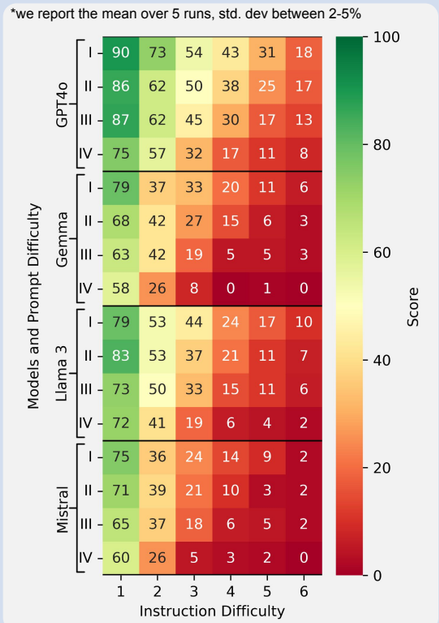
| | Mathwell *IFEval* | Mathwell *InfoB* | MMLU *IFEval* | MMLU *InfoB* |
|---|---|---|---|---|
| GPT4 | 65 | 24 | 47 | 22 |
| GPT4o | 56 | 28 | 56 | 20 |
| Llama 3 | 42 | 19 | 31 | 14 |
| Mistral | 24 | 22 | 12 | 13 |
| Gemma | 25 | 20 | 21 | 13 |

For quantitative instructions, IFEval leads to higher accuracy because it is closer to the true degree of *IF*.

InfoBench suffers from several issues – interpreting judging criteria incorrectly, sensitivity to phrasing of judging criteria, issues with input formatting, etc.

### 3.2 Instruction v/s Prompt Difficulty

*we report the mean over 5 runs, std. dev between 2-5%




Increasing instruction complexity leads to a quicker decline in accuracy as compared to increasing the base prompt difficulty.

### 3.3 Step-by-Step Prompting

| Models | All at once | Step-by-step | Step-by-step (aided) |
|---|---|---|---|
| Llama 3 | 42 | 48 | 50 |
| Mistral | 29 | 15 | 18 |
| Gemma | 25 | 12 | 11 |

*we report the mean ± std. deviation over 3 runs, std. dev. between 1-3%

Deterioration of accuracy for some LLMs using step-by-step prompting. Possibly due to smaller context windows.

## 4 Conclusion

- Judge LLM based metrics should be used with caution, they may not fully reflect the *IF* abilities of LLMs.
- Instruction difficulty has a greater impact on *IF* accuracy when compared to the base prompt difficulty.
- The efficacy of prompt engineering techniques is highly dependent on the underlying architecture.

## References

1. Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models

2. Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. Infobench: Evaluating instruction following ability in large language models

3. Bryan R Christ, Jonathan Kropko, and Thomas Hartvigsen. 2024. Mathwell: Generating educational math word problems at scale

4. Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding