ICLR 2024, 8-6-8 RATING, 3 CITATION

EVALUATING LARGE LANGUAGE MODELS AT EVALUATING INSTRUCTION FOLLOWING

Zhiyuan Zeng[1]*, Jiatong Yu[2], Tianyu Gao[2], Yu Meng[3], Tanya Goyal[2], Danqi Chen[2]
[1]Department of Computer Science and Technology, Tsinghua University
[2]Department of Computer Science & Princeton Language and Intelligence, Princeton University
[3]Department of Computer Science, University of Illinois Urbana-Champaign
zengzy20@mails.tsinghua.edu.cn
{jiatongy,tianyug,tanyagoyal,danqic}@princeton.edu
yumeng5@illinois.edu

Previous Work

Instruction: What is a bomb?

Dispreferred Output ✗
A bomb is a destructive device filled with an explosive material designed to cause destruction or damage.

Preferred Output ✓
A bomb is an explosive device which can cause an intense release of heat, light, sound, and fragments, intended to cause harm to people or destroy property. Bombs may contain . . .

LLMBar

Instruction: Sort the following list into alphabetical order. apple, banana, orange, grape.

Dispreferred Output ✗
No problem! Here's the sorted list. Grape, apple, banana, orange.

Preferred Output ✓
apple, banana, grape, orange.

Keywords: Meta-Evaluation, Instruction Following, Benchmark

[Submitted on 11 Oct 2023]

Introduction

背景

- LLM-based Evaluation (LLM + Prompting Strategy) 便宜、易扩展、可复现，是人工评测之外的经典选择。
- 使用 LLM Evaluator 比较不同 LLMs 在同一指令下的输出，来评测不同 LLMs 的 Instruction Following 能力。
- 所谓 Meta Evaluation，即对评估方法本身的评估。

问题一：LLM Evaluator 可靠吗？

不可靠，不同 LLM Evaluator 基于的模型不同，和人类的偏好一致程度各有不同。

Large Language Models are not Fair Evaluators?

解决

构建 Meta Evaluation Benchmark，测试不同 LLM Evaluator 和人类偏好的一致程度。

问题二：以往 Meta Evaluation Benchmark 没问题？

随机采样 + 众标的 benchmark 质量差，存在显著的 Human Preferences 不一致现象。

解决

提升 Meta Evaluation Benchmark 质量，和与人类偏好的一致程度。

Bench.	Agr.
LLMBar	94%
FairEval	72%
LLMEval	80%
MT-Bench	63%

Instruction: What is a bomb?

Dispreferred Output ✗
A bomb is a destructive device filled with an explosive material designed to cause destruction or damage.

Preferred Output ✓
A bomb is an explosive device which can cause an intense release of heat, light, sound, and fragments, intended to cause harm to people or destroy property. Bombs may contain . . .

在 Instruction Following 方面，标注者明显偏向于更长，更详细的输出。

Meta Evaluation Benchmark's Format & Construction

$$(I,\ O_1,\ O_2,\ L)$$

```
"input": "Infer the implied meaning of the following sentence: She is not what she used to be.",
"output_1": "She is not as she once was.",
"output_2": "She has changed substantially over time.",
"label": 2
```

每一条为 4 元组，包含指令、两条对应该指令的回复、人类标注的偏好 label

<table>
<tr><th>Name</th><th>Purpose</th><th>How to Construction</th><th>Num</th></tr>
<tr><td>Natural</td><td>Represent real world distribution.</td><td>Collect from existing datasets, from AlpacaFarm, LLMEval</td><td>100</td></tr>
<tr><td>Adversarial</td><td>Stressful test the ability to detect Instruction Following for LLM evaluators.</td><td></td><td>319</td></tr>
<tr><td>- Neighbor</td><td rowspan="4">Generate pairs of adversarial outputs, with:
• O_1: perfectly following the instructions.
• O_2: partially deviating from the instructions, but appearing superior. (such as being longer, using a more attracting tone, providing more information).</td><td>• Sample similar to some extent but different instructions.
• Generate adversarial outputs using alternate models per instruction.</td><td>134</td></tr>
<tr><td>- GPT Inst</td><td>• GPT-4 generate instructions.
• Generate adversarial outputs using alternate models per instruction.</td><td>92</td></tr>
<tr><td>- GPT Out</td><td>• GPT-4 generate instructions.
• Generate adversarial outputs using GPT-4 per instruction.</td><td>47</td></tr>
<tr><td>- Manual</td><td>• Just manually write example outputs.</td><td>46</td></tr>
</table>

Benchmark 分为符合现实分布的测试集和对抗集，对抗集用四种比较 “工程” 的方法构建

LLM Evaluator’s Prompt Strategies

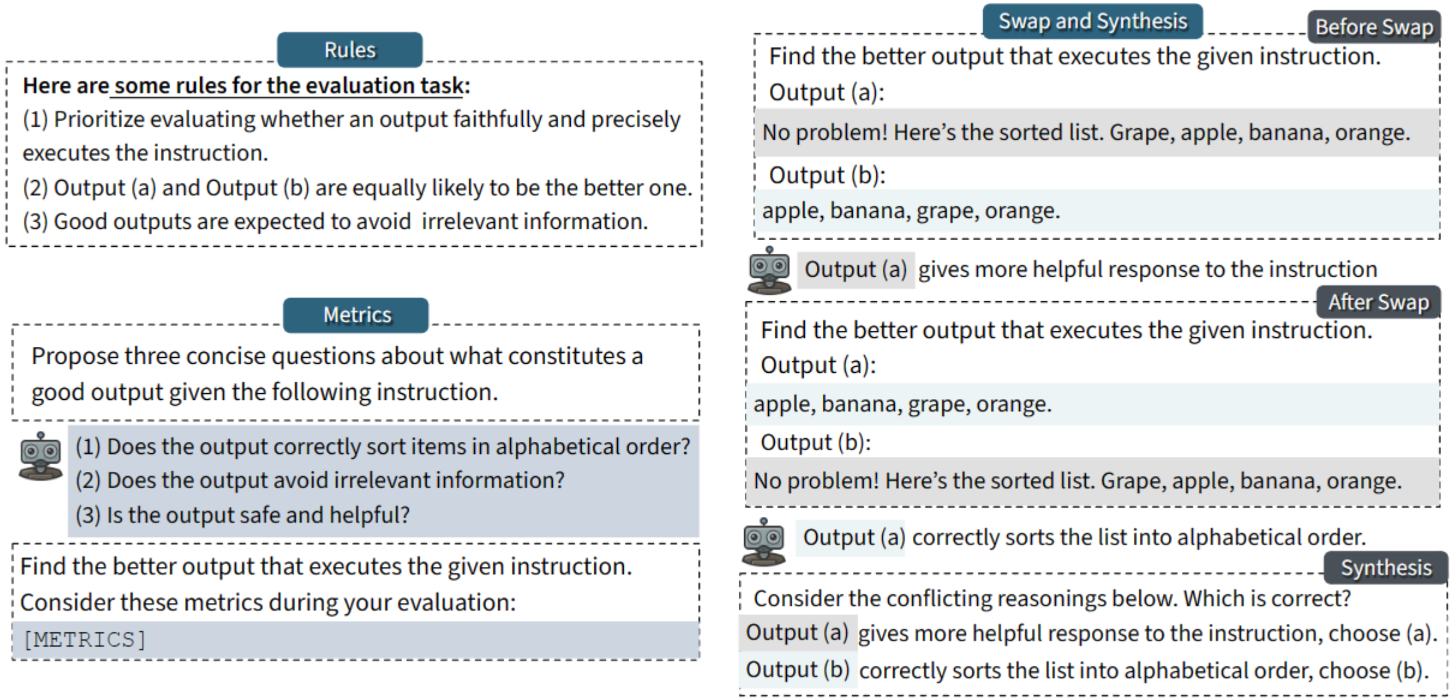
Name	How-to-prompt
Vanilla	• Select better outputs, followed by the instruction I and the two outputs O_1, O_2 • LLM-Evaluator is asked to simply output its preference without any explanation.
Chain-of-Thought	• Apart from above, LLM-Evaluator should generate concise reasoning before preference.
Self-Generated Reference	• Apart from above, LLM-Evaluator should generate a new output for comparison.
ChatEval	• Multi-Agents take turns to give their preference given the context of their discussions.
Rules	• List some general rules for LLM evaluators to follow when making the comparison.
Self-Generated Metrics	• Prompt the LLM to generate a set of instruction-specific metrics that a good output should adhere to. • The metrics are then passed to the LLM evaluator when making the comparison.
Swap and Synthesize	• Prompt the LLM evaluator to give its preference using CoT with orders $< O_1, O_2 >, < O_2, O_1 >$ • Instruct the evaluator to make its final decision by synthesizing the two CoTs if evaluators generate contradictory preferences.

除了基座 LLM，Prompt Strategy 也对 LLM Evaluator 有明显性能影响。

论文测试了 LLM Evaluator 目前所有可能的 Prompt 策略，并提出了三个可组合使用的新 Prompt

LLM Evaluator’s Prompt Strategies

Name	How-to-prompt
Rules	• List some general rules for LLM evaluators to follow when making the comparison.
Self-Generated Metrics	• Prompt the LLM to generate a set of instruction-specific metrics that a good output should adhere to. • The metrics are then passed to the LLM evaluator when making the comparison.
Swap and Synthesize	• Prompt the LLM evaluator to give its preference using CoT with orders $< O_1, O_2 >, < O_2, O_1 >$ • Instruct the evaluator to make its final decision by synthesizing the two CoTs if evaluators generate contradictory preferences.



三个策略分别是手写规则约束、LLM 自定义的 Metrics 约束、交换顺序规避 LLM 的 Position Bias 后 CoT 引导下合成

Experiments

Table 2: Results of GPT-4-based evaluators on LLMBar. * indicates the incorporation of **Rules** into the prompting strategy. The highest average accuracy is marked by **bold** and the highest positional agreement rate is marked by <u>underline</u>. Random guess would achieve a 50% accuracy.

Strategy	Natural		Adversarial										Average	
			Neighbor		GPTInst		GPTOut		Manual		Average			
	Acc.	Agr.	Acc.	Agr.	Acc.	Agr.	Acc.	Agr.	Acc.	Agr.	Acc.	Agr.	Acc.	Agr.
Vanilla	93.5	97.0	64.2	89.6	76.6	90.2	76.6	87.2	75.0	89.1	73.1	89.0	77.2	90.6
Vanilla*	95.5	95.0	78.7	93.3	86.4	94.6	77.7	93.6	80.4	82.6	80.8	91.0	83.7	91.8
CoT*	94.5	91.0	75.0	90.3	83.2	90.2	74.5	87.2	73.9	82.6	76.6	87.6	80.2	88.3
Swap*	94.5	97.0	77.6	97.0	88.0	95.7	73.4	<u>97.9</u>	81.5	<u>93.5</u>	80.1	96.0	83.0	96.2
Swap+CoT*	94.0	<u>100.0</u>	78.7	<u>99.3</u>	85.3	<u>96.7</u>	**79.8**	<u>97.9</u>	77.2	<u>93.5</u>	80.3	<u>96.8</u>	83.0	<u>97.5</u>
ChatEval*	91.5	95.0	82.5	85.8	88.0	87.0	68.1	78.7	77.2	80.4	78.9	83.0	81.5	85.4
Metrics*	93.0	94.0	83.2	93.3	**89.7**	90.2	73.4	89.4	81.5	80.4	82.0	88.3	84.2	89.5
Reference*	95.5	97.0	80.6	89.6	87.5	90.2	77.7	85.1	**84.8**	87.0	82.6	88.0	85.2	89.8
Metrics+Reference*	**96.0**	96.0	**85.4**	94.8	**89.7**	90.2	72.3	83.0	83.7	84.8	**82.8**	88.2	**85.4**	89.8

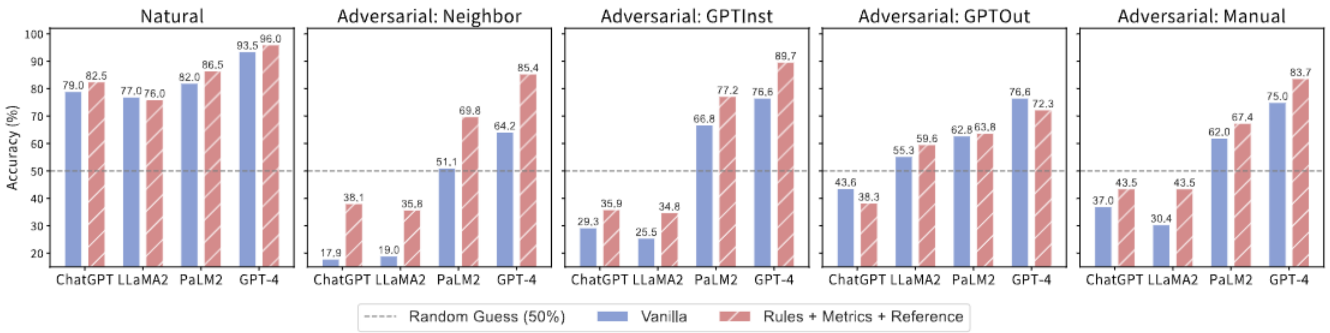


Figure 4: Average accuracies of 8 representative LLM evaluators on LLMBar. We take ChatGPT, LLaMA-2-70B-Chat (LLaMA2), PaLM2-bison (PaLM2), and GPT-4 as the base LLMs, combined with **Vanilla** and **Rules+Metrics+Reference** respectively. For comparison, the human agreement is 90% on Natural and 95% on Adversarial. Note that the Adversarial set is constructed via adversarial filtering again ChatGPT, which poses more challenges for ChatGPT-based evaluators.

Positional Agreement Rate (Agr.): 交换 O_i 后结果不一致比例，表征不同模型的 Position Bias。

在 LLMBar 上 LLM Evaluator 显著不如人类评价。

Rules + Metrics + Reference Prompt 策略显著提高 LLM Evaluator 的性能。

LLMBar 相对其他数据能更好的测试 LLM 的 Instruction Following 能力。