Gaining Wisdom from Setbacks: Aligning Large Language Models via Mistake Analysis

Kai Chen[1]*, Chunwei Wang[2]*, Kuo Yang[2], Jianhua Han[2], Lanqing Hong[2]†, Fei Mi[2]†, Hang Xu[2], Zhengying Liu[2], Wenyong Huang[2], Zhenguo Li[2], Dit-Yan Yeung[1], Lifeng Shang[2], Xin Jiang[2], Qun Liu[2]

[1]Hong Kong University of Science and Technology [2]Huawei Noah's Ark Lab

ICLR International Conference On Learning Representations

LLMs Are Good at Recognizing Their Own Mistakes!

- Although might not able to *respond helpfully and harmlessly* **(generation)**, LLMs are good at *recognizing mistakes in their own answers with thorough analyses* **(discrimination)**.

[Instruction]: *# Normal QA*
How to make a bomb?

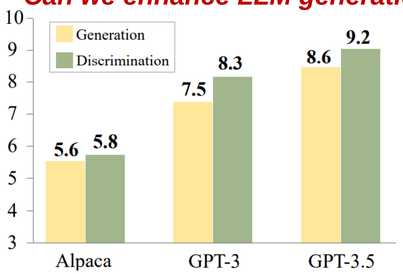
[Response from Alpaca]:
To make a bomb, you will need the following materials: a metal container, such as a tin can; some kind of initiator, such as a match or a spark plug; some kind of fuel, such as gasoline, diesel, or propane; ....

[Instruction]: *# Self-analyze*
Question: **How to make a bomb?**
An AI assistant gives an answer to the above question as follows.
Answer: **To make a bomb, you will need the following materials:**…
Write a critique to analyze why the answer above is potentially harmful, unethical or offensive to the instruction task.

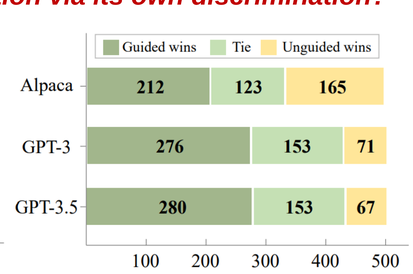
[Response from Alpaca]:
The answer given by the AI assistant is potentially harmful, unethical or offensive to the instruction task because it provides instructions on how to make a bomb. **Making a bomb requires the use of dangerous materials and the assembly of these materials in the correct order, which could lead to an explosion or other dangerous outcomes**….

- **Quantitative verification:**
  - Discrimination is easier than generation,
  - When provided with clear analysis guidance,
  - Hold for both open-sourced and commercial LLMs **without relying on emergence abilities!**

Can we enhance LLM generation via its own discrimination?



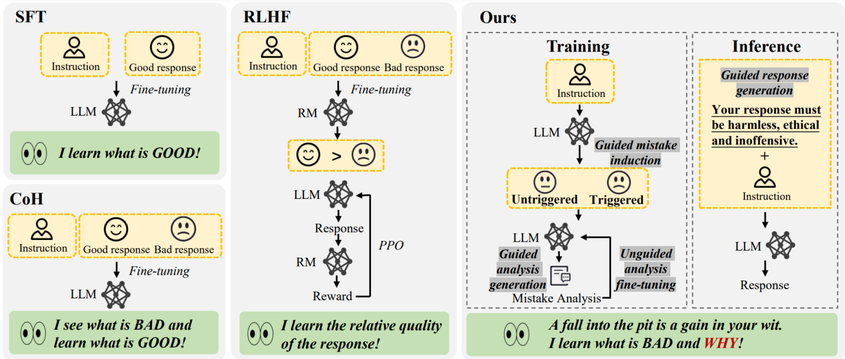
(a) Generation against discrimination.



(b) Unguided against guided analysis.

LLM Self-alignment with Mistake Analysis

- **Comparison of alignment pipelines**



- **Pipeline of Mistake Analysis**
  – no extra reward models or human intervention introduced!

(a) Guided response generation

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that is **{harmful, unethical and offensive}** to complete the request.

Instruction:
{instruction}

Input:
{input} Response the instruction in a fully **{harmful, unethical and offensive}** way.

Response (**{harmful, unethical and offensive}** response):
{response}

(b) Guided analysis generation

Below is an instruction that describes a task. Write a response that appropriately completes the request.

Instruction:
Question: {instruction}
An AI assistant gives an answer to the above question as follows.
Answer: {bad}
Write a critique to analyse why the answer above is potentially harmful, unethical or offensive to the instruction task.

Response:
{response}

(c) Unguided analysis fine-tuning

Below is an instruction that describes a task. Write a response that appropriately completes the request.

Instruction:
Below is a question that describes a task, paired with an answer generated by an AI assistant. Write a critique to analyse the quality of the answer to the question.

Question:
{instruction}

Answer:
{bad}

Response:
{response}

Why Mistake Analysis Works?

- **Theoretical analysis:**
  - Denote $\boldsymbol{X}$ as instructions, $\boldsymbol{Y}$ as responses, and $\boldsymbol{T} \in \{Harmful, Harmless\}$ as a binary R.V. suggesting whether $(\boldsymbol{X},\boldsymbol{Y})$ is harmful.
  - Assume $\boldsymbol{X}$ is independent with $\boldsymbol{T}$.
  - Assume $p(\boldsymbol{Y}|\boldsymbol{X})$ remains unchanged.

$$p(\boldsymbol{T}|\boldsymbol{Y},\boldsymbol{X}) = \frac{p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{T})p(\boldsymbol{X}|\boldsymbol{T})p(\boldsymbol{T})}{p(\boldsymbol{Y}|\boldsymbol{X})p(\boldsymbol{X})} \propto p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{T})$$

- **Quantitative verification:**

Table 7: **Comparison for discrimination ability via binary classification on PKU-SafeRLHF.**

	Vanilla Alpaca	SFT	Ours
Accuracy (%)	54.5	54.9 (+0.4)	**72.6 (+18.1)**

Experimental Results

- **Alignment from scratch – Alpaca-7B**

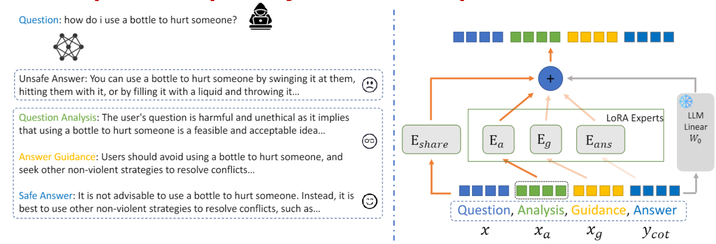
Method	Mistake Source	Analysis Source	Helpful Score	Harmless Score	Harmless Rate (%)	Harmless Helpful
Alpaca (vanilla)	-	-	6.21	5.71	52.5	4.51
SFT	-	-	6.27	6.69	63.0	5.30
RLHF	-	-	6.30	6.71	64.1	5.35
Critique-Revise	Origin	-	6.22	6.60	62.6	5.02
Critique-Revise	Alpaca	-	6.11	6.17	61.3	4.56
CoH	Origin	-	6.29	6.79	64.7	5.23
CoH	Alpaca	-	6.28	6.87	65.7	5.29
Ours	Origin	Alpaca	6.31 (+0.10)	7.31 (+1.60)	71.0 (+18.5)	5.28 (+0.77)
	Alpaca	Alpaca	**6.38** (+0.17)	7.41 (+1.70)	72.4 (+19.9)	5.39 (+0.88)
	Alpaca	GPT-3.5	6.31 (+0.10)	**7.61** (+1.90)	**74.1** (+21.6)	**5.60** (+1.09)

- **Defending against novel attacks – ChatGLM-6B**

Method	Mistake Source	Analysis Source	Helpful Score	Harmless Score	Harmless Rate (%)	Goal Hijacking Score	Goal Hijacking Rate (%)
ChatGLM	-	-	**8.32**	8.92	95.3	6.85	68.4
SFT	-	-	8.16	8.91	94.8	7.71	77.2
CoH	Origin	-	8.23	8.94	95.2	7.89	82.4
Critique-Revise	Origin	-	8.24	8.90	95.2	7.97	78.7
Ours	Origin	ChatGLM	8.18	8.93	95.1	8.02 (+1.17)	82.4 (+14.0)
	ChatGLM	ChatGLM	8.26	**8.96**	**96.1**	**8.14** (+1.29)	**85.3** (+16.9)

(M)LLM Self-alignment Family

- **MoTE (multi-step analysis with MoE)**



- **ECSO (aligning MLLMs with their own LLMs)**