



山东大学  
SHANDONG UNIVERSITY



Universiteit  
Leiden

# Improving the Robustness of Large Language Models via Consistency Alignment

---

Yukun Zhao, Lingyong Yan, Weiwei Sun, GuoLiang Xing,  
Chong Meng, Shuaiqiang Wang, Zhicong Cheng,  
Zhaochun Ren, Dawei Yin



# Outline

---

- Introduction
- Robustness on Instruction Following
- Training Framework
- Experiments
- Conclusion
- Limitations



# Introduction

---

- Large Language Models (LLMs) demonstrated remarkable capabilities,
  - Understanding human instructions and generating helpful responses.
  
- The robustness of current LLMs is still far from promising, i.e., the inconsistency problem.



# Introduction

...  
Catherine Willows: Okay, no phone, no friends, no nothing.  
Lindsey Willows: For how long?  
Catherine Willows: A month  
Lindsey Willows: Whatever  
Catherine Willows: Hey, you want to make it two ?  
...



... Use your language skills to **determine what the element being referred to by the underlined number.** Like number ...



two



... Employ your knowledge to **determine the referent of the highlighted number.** The numbers will be marked with two underlines surrounding like number ...



The referent of the highlighted number "two" is months

- ❑ LLMs generate inconsistent answers for the identical tasks.
- ❑ Two GPT-4 examples.
- ❑ Hindering practical applications.



# Introduction

---

- Recent work explored the prompt-sensitive problem.
- They proposed to optimize the task instruction to elicit the best performance.
- We quantitatively analyze the consistency of the LLMs' generation.
- We then propose a novel training framework via consistency alignment.



# Inconsistency in LLMs

---

## □ Consistency Definition:

$$\mathcal{R} = \mathbb{E}_{q_i, q_j \in Q} \left[ \mathbb{E}_{y_i \sim Y(q_i), y_j \sim Y(q_j)} [\mathbf{sim}(y_i, y_j)] \right]$$

- $Q$ , all conceivable linguistic paraphrases;  $Y$ , possible responses

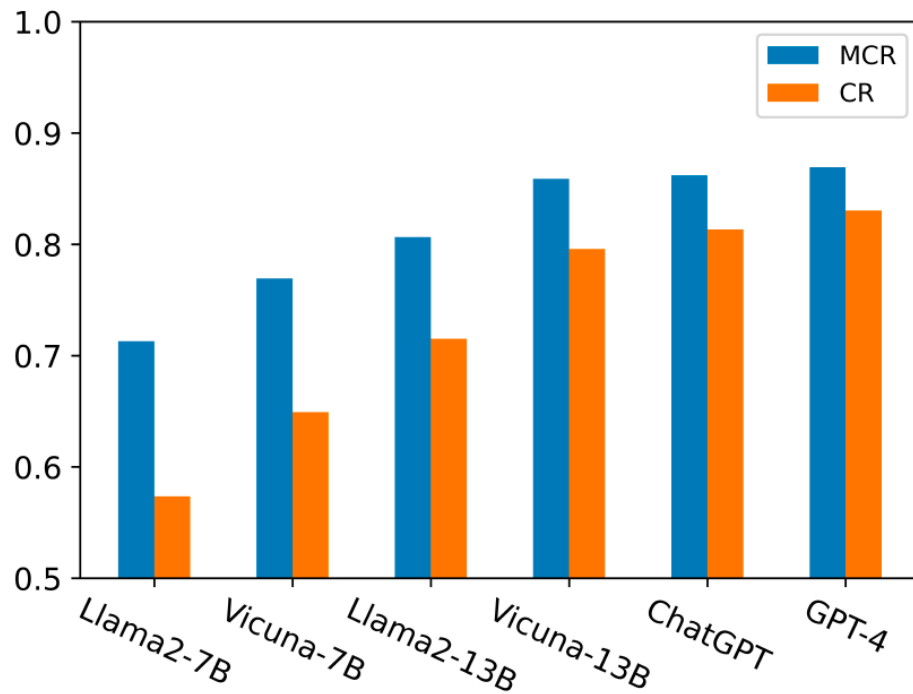
## □ Consistency metrics: consistency rate $CR$ , and maximum consistency rate $MCR$

$$CR = \frac{1}{|Q|} \sum_{Q_k \in Q} \sum_{y_i \in Y_k} \sum_{y_j \in Y_k, j \neq i} \frac{\mathit{sim}(y_i, y_j)}{\binom{|Y_k|}{2}} \quad \mathit{sim}(y_i, y_j) \in \{0,1\}$$

$$MCR = \frac{1}{|Q|} \sum_{Q_k \in Q} \frac{|\Omega_k^{max}|}{|Y_k|} \quad \Omega_k \text{ is a cluster of consistent responses}$$



# Inconsistency in LLMs



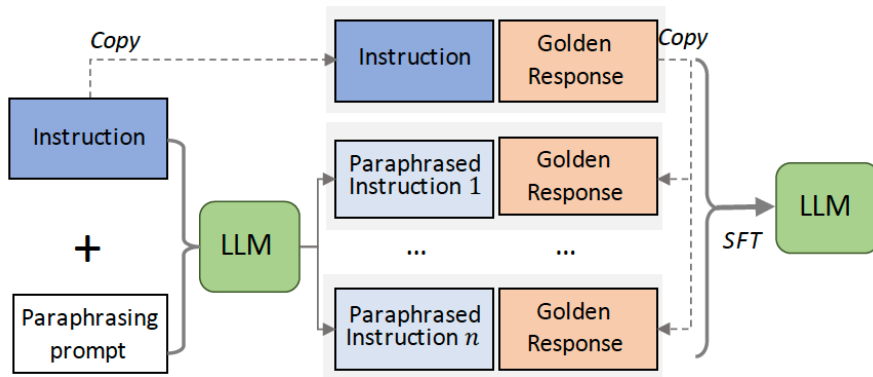
- Consistency metrics of current LLMs on Super Natural Instructions.
- Necessity to improve the robustness especially the smaller one



# Training Framework

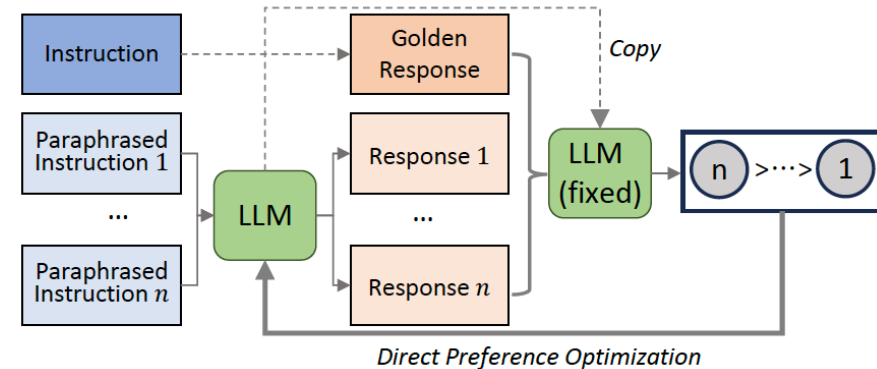
## Supervised fine-tuning with instruction augmentation

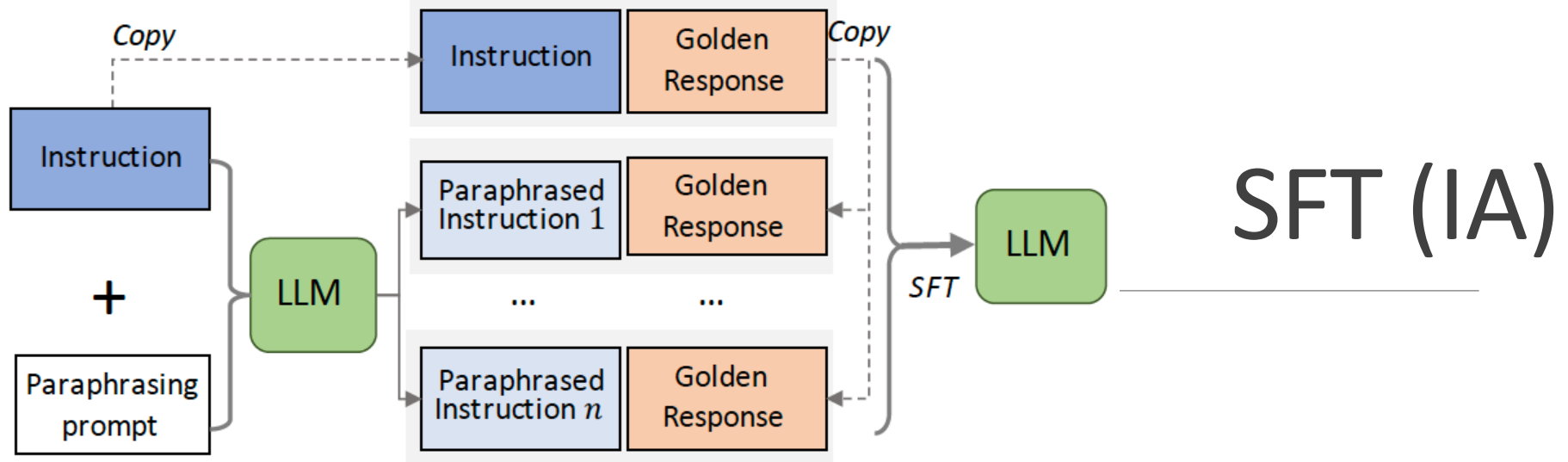
- SFT (IA) for short



## Consistency alignment training with automatic feedback

- CAT for short





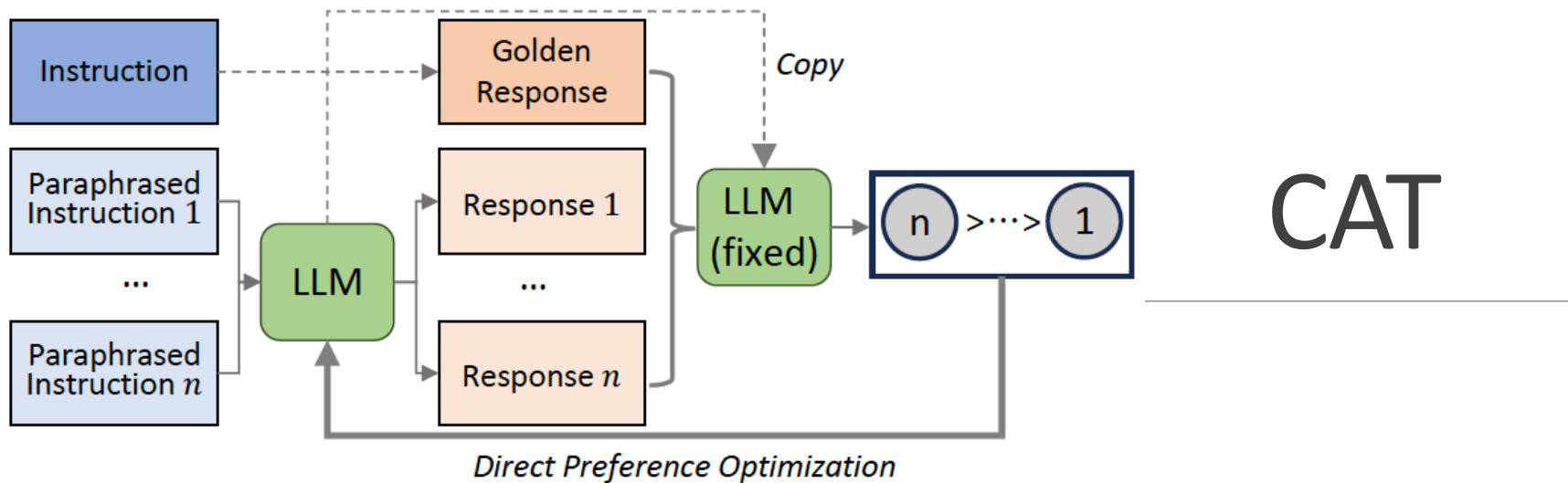
## □ Instruction Augmentation:

- Prompt LLMs to paraphrase original instructions into several re-phrasings.
- Prompt: *Paraphrase the input to have different words and expressions but have the same meaning as the original sentences ...*

## □ Supervised Fine-tuning:

- Training Set  $S = \bigcup_k \bigcup_j^n \bigcup_i^m \{a_j^k, x_i^k, y_i^k\}$ .  $k$  tasks,  $n$  instructions,  $m$  instances.
- Loss: 
$$L_{sft} = - \sum_t \log P(y_{i,t} | a, x_i, y_{i < t})$$





# CAT

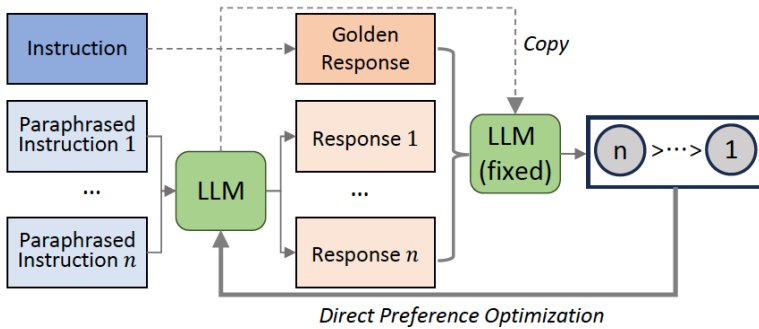
## Collecting training pairs:

- For input  $x_i$ , utilize the trained model to generate  $n$  responses for  $n$  instructions

## Self rewards:

- Prompt the trained model to give a reward  $r_i$  for each response  $y_i$
- Sub-reward: Expected answer type  $r_i^T \in \{0, 1\}$ , with prompt:
  - *Determine whether the answer is the expected answer type for the question Q ...*
- Sub-reward: Correctness  $r_i^C \in \{0, 1\}$ , with prompt:
  - *Determine whether the answer is Correct or Incorrect for the question Q ...*





# CAT

## Self rewards:

$$r_i = \begin{cases} 0, & r_i^T = 0 \\ 1, & r_i^T = 1 \wedge r_i^C = 0 \\ 2, & r_i^T = 1 \wedge r_i^C = 1 \end{cases}$$

## Training Objective:

- Training pairs: each input  $x_i, U_{i,j} < y_j, y_i >$  where  $r_j > r_i \wedge r_j = 2$

- Loss:

$$L_{rank} = \sum_{r_i < r_j} \max(0, p_i - p_j)$$

$$L = L_{rank} + \lambda * L_{sft}$$



# Experiments

---

- Experimental setup
- Main results
- Detailed Analysis
- Human Evaluation



# Experimental setup

---

- Dataset: Super Natural Instructions. Construct test set I/II.
- Models: Vicuna-7B, Vicuna-13B, LLama 2-7B and Llama 2-13B
- Baselines: SFT, off-the-shelf LLMs ChatGPT and GPT-4
- Evaluation Metrics:
  - Robustness metrics  $CR$ , and  $MCR$  on test set I ;
  - Correctness metrics ROUGE-1 and ROUGE-L on test set I/II



# Experimental setup

---

## □ Implementation Details:

- Use Vicuna-7B, Vicuna-13B, and ChatGPT to paraphrase the original task instructions.
- Randomly sample 10 instructions \* 10 instances for each task.
- Training SFT (IA) using FastChat, with epochs=3, lr=2e-5, etc.
- Self-wards are from LLMs themselves, fine-tuned Vicuna-7B, Vicuna-13B, Llama 2-7B, and Llama 2-13B
- CAT, using revised Llama-efficient DPO, with epochs=3, lr=1e-5,  $\lambda=1$ , etc. with LoRA.



# Main Results on test set I

	CR	MCR	ROUGE-1	ROUGE-L
GPT-4	0.8303	0.8693	0.3870	0.3751
ChatGPT	0.8134	0.8620	0.3022	0.2744
Vicuna-7B	0.6492	0.7694	0.1385	0.1266
Vicuna-7B + SFT	0.7092	0.8123	0.3782	0.3672
Vicuna-7B + SFT (IA)	0.7753	0.8504	0.3894	0.3757
Vicuna-7B + SFT (IA) + CAT	0.8298	0.8743	0.4187	0.4097
Vicuna-13B	0.7959	0.8589	0.1724	0.1596
Vicuna-13B + SFT	0.8017	0.8490	0.4028	0.3903
Vicuna-13B + SFT (IA)	0.8267	0.8619	0.4131	0.4014
Vicuna-13B + SFT (IA) + CAT	0.8390	0.8804	0.4276	0.4185
Llama2-7B	0.5735	0.7129	0.0637	0.0492
Llama2-7B + SFT	0.7702	0.8308	0.2682	0.2560
Llama2-7B + SFT (IA)	0.7921	0.8475	0.2901	0.2733
Llama2-7B + SFT (IA) + CAT	0.8107	0.8521	0.3012	0.2806
Llama2-13B	0.7151	0.8065	0.0737	0.0627
Llama2-13B + SFT	0.7505	0.8180	0.3085	0.2975
Llama2-13B + SFT (IA)	0.7589	0.8282	0.3379	0.3280
Llama2-13B + SFT (IA) + CAT	0.8100	0.8601	0.3711	0.3502

## Observation:

- + SFT > Vanilla
- + SFT (IA) > +SFT
- + SFT (IA) + CAT > + SFT (IA)

## Effectiveness



# Main Results on test set II

	ROUGE-1	ROUGE-L
GPT-4	0.4506	0.4408
ChatGPT	0.3187	0.3051
Vicuna-7B	0.1702	0.1570
+SFT	0.4085	0.3929
+SFT (IA)	0.4122	0.3984
+SFT (IA) + CAT	0.4391	0.4285
Vicuna-13B	0.2102	0.1972
+SFT	0.4234	0.4071
+SFT (IA)	0.4477	0.4350
+SFT (IA) + CAT	0.4683	0.4417
Llama2-7B	0.0684	0.0513
+SFT	0.2743	0.2614
+SFT (IA)	0.3163	0.2903
+SFT (IA) + CAT	0.3189	0.2977
Llama2-7B	0.0745	0.0643
+SFT	0.3351	0.3215
+SFT (IA)	0.3697	0.3587
+SFT (IA) + CAT	0.4289	0.4066

□ Comparison on the standard test set.

□ Observation:

- + SFT > Vanilla
- + SFT (IA) > +SFT
- + SFT (IA) + CAT > + SFT (IA)



# Detailed Analysis

---

## □ The choice of Rewards:

- CAT using  $r_i^C$  only or using  $r_i^T + r_i^C$
- CAT rewards from fine-tuned version or the vanilla Vicuna-13B
- Report the ROUGE values on test set I
- Observation:  $r_i^T + r_i^C > r_i^C$ ; A strong LLM is better for rewarding

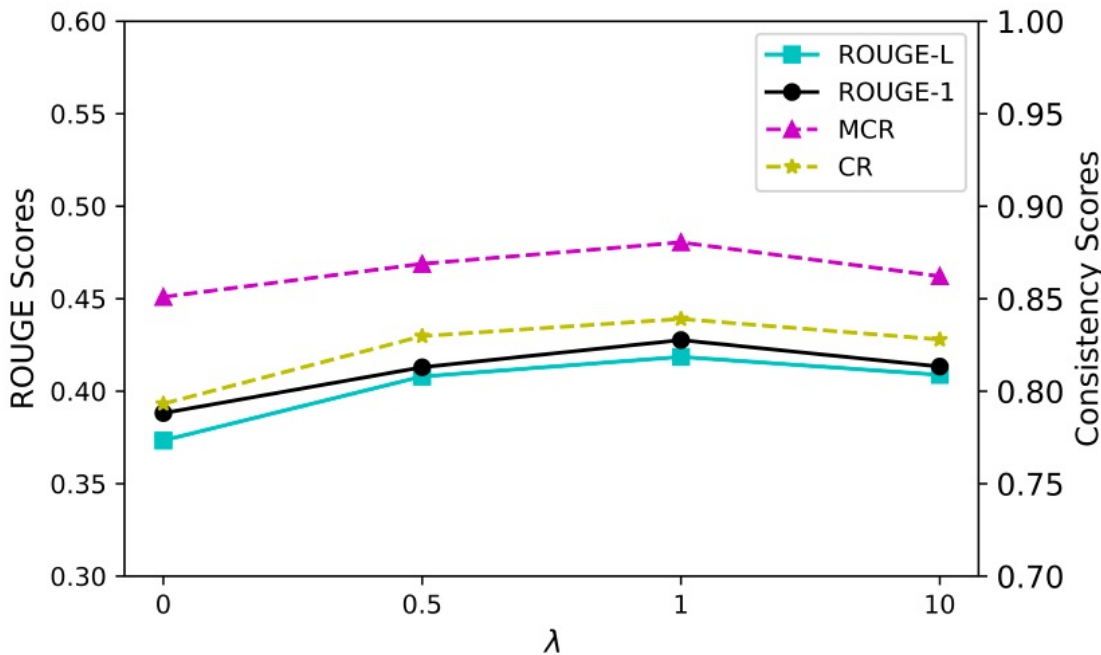
Rewards	ROUGE-1	ROUGE-L
$r^C$ from SFT	0.4123	0.4051
$r^C + r^T$ from SFT	0.4276	0.4185
$r^C + r^T$ from Vicuna-13B	0.3962	0.3877



# Detailed Analysis

## □ The choice of $\lambda$

- The performance of diff.  $\lambda$  in the loss.



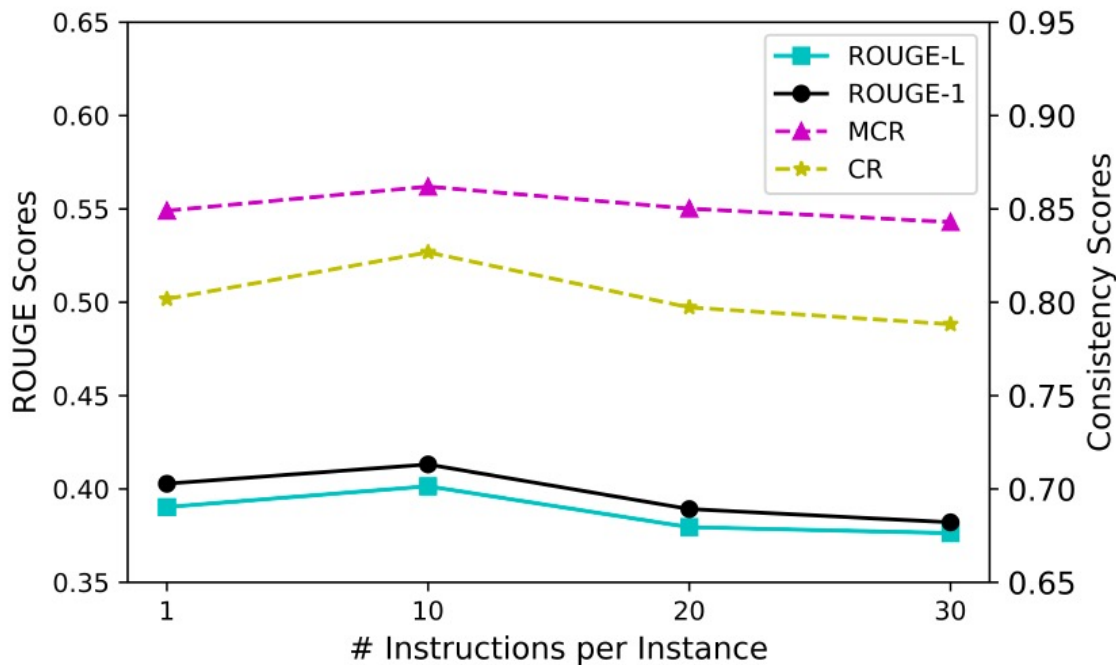
## Observations:

- The necessity of adding the SFT loss
- The necessity of CAT, learning from negative generations.



# Detailed Analysis

- The performance of diff. number of augmented instructions.
  - The size of training set and other hyperparameters are fixed.



## Observations:

- The necessity of augmenting instr..
- The necessity of sufficiently trained instances.



# Human Evaluation

---

## □ Human evaluation on different trained models.

- Report the diff. ratios, and evaluate the diff. responses with 0/1/2 and report wins, ties, and losses.

Strategy	Baseline	diff.	win	tie	lose
CAT+SFT(IA)	Vanilla	83	48	44	8
CAT+SFT(IA)	SFT	46	32	55	13
CAT+SFT(IA)	SFT(IA)	27	31	60	9

## • Observations:

- Superior performance when use CAT + SFT(IA).
- A model can be continually improved with additional CAT.



# Conclusion

---

- We investigate the robustness of current LLMs in terms of the consistency of the generated responses.
- We introduce a novel training framework, i.e, SFT(IA) + CAT , which helps to boost the robustness of LLMs.
- The method is self rewarded, without the need for additional human guidance or external reward models.
- We conduct extensive experiments to verify the effectiveness of the proposed training framework and each component.



# Limitations

---

- ❑ The self-rewarding may be limited when it is applied in a less aligned LLM.
- ❑ The diversity of the verbalized instructions may be limited compared with end-users.





山东大学  
SHANDONG UNIVERSITY



Universiteit  
Leiden

# Thanks, Q&A

---

YUKUN ZHAO@LREC-COLING 2024

YUKUNZHAO.SDU@GMAIL.COM

