

Improving the Robustness of Large Language Models via Consistency Alignment

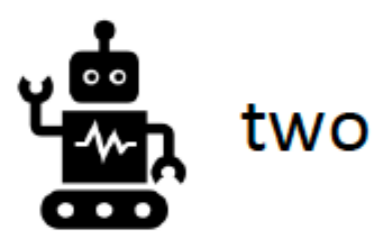
Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren[#], Dawei Yin[#]

Current Large Language Models (LLMs) generate inconsistent answers for the identical tasks.

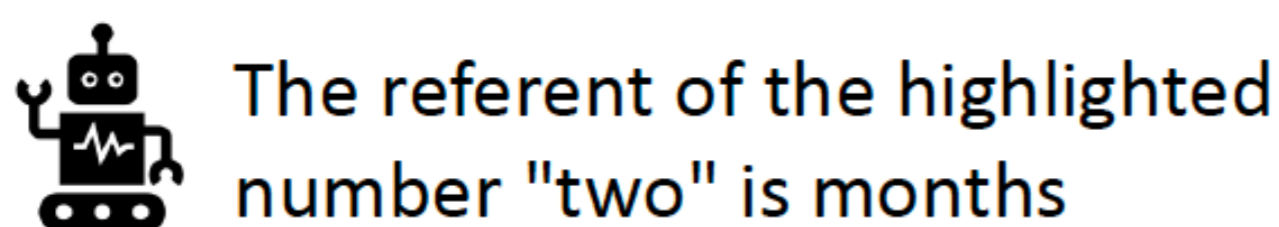
...
Catherine Willows: Okay, no phone, no friends, no nothing.
Lindsey Willows: For how long?
Catherine Willows: A month
Lindsey Willows: Whatever
Catherine Willows: Hey, you want to make it two ?
...

- GPT-4 generate inconsistent responses for *the referent of the number* task.
- Hindering practical applications.

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



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



Inconsistency in LLMs

Consistency Definition:

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

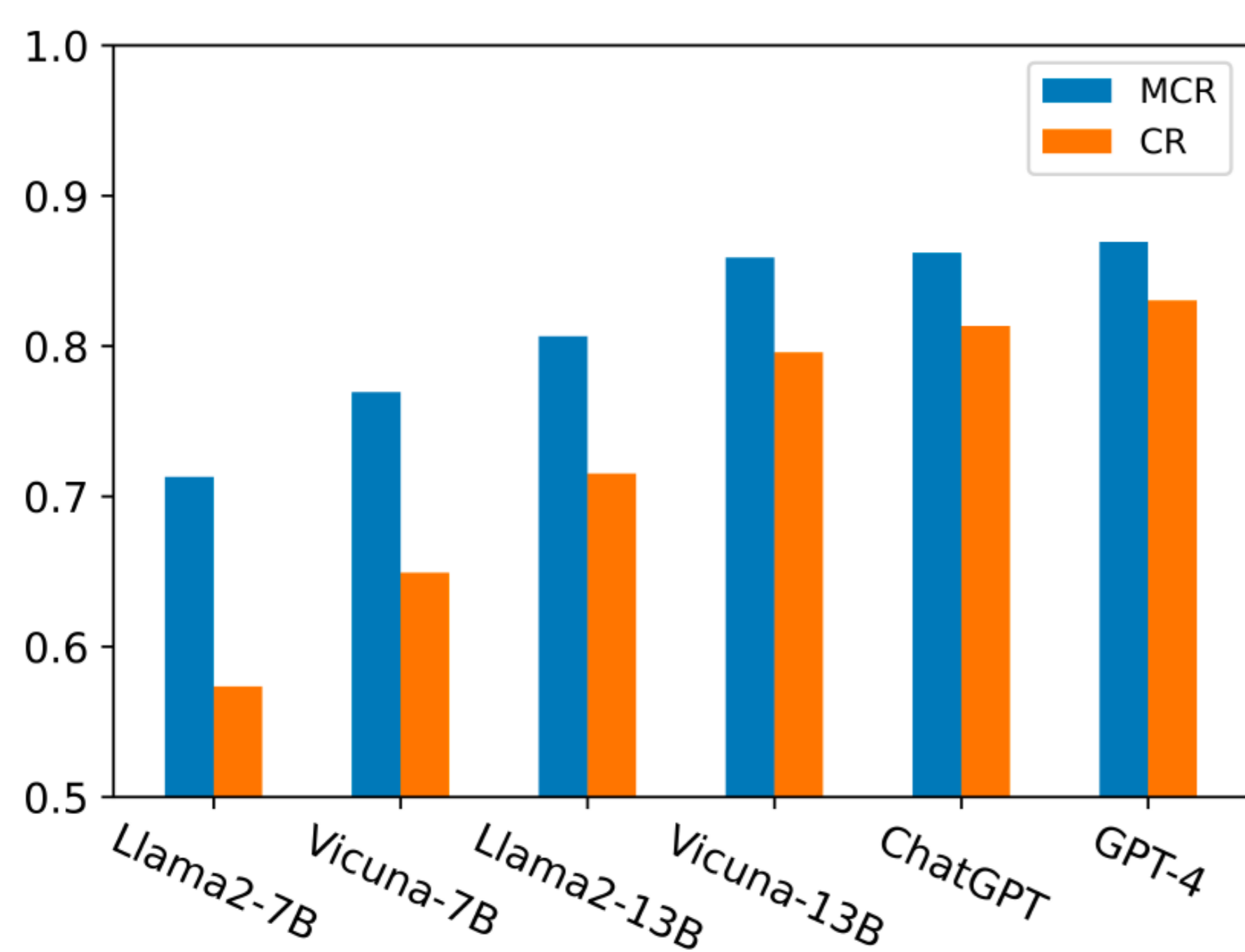
Q , all conceivable linguistic paraphrases; Y , possible responses

- **Consistency metrics:** consistency rate CR , and maximum consistency rate MCR

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

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

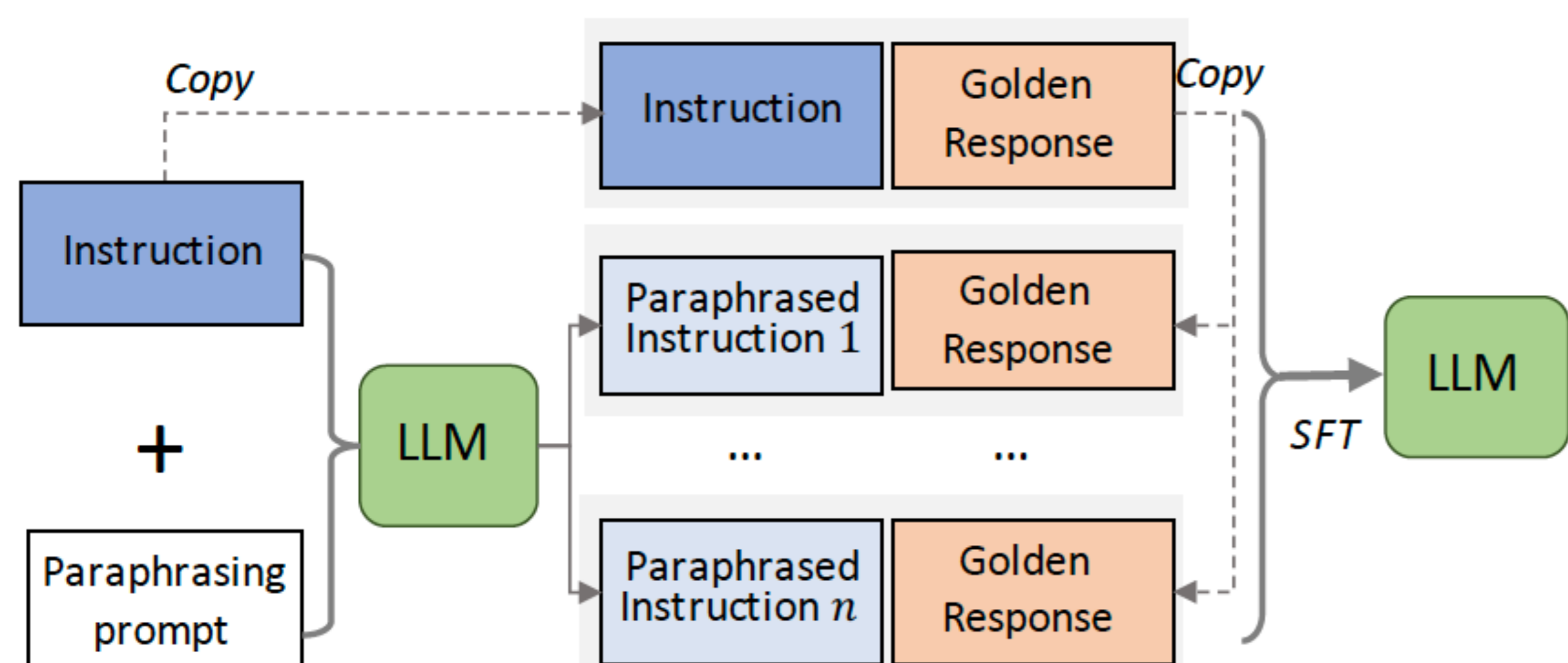
Current Consistency of LLMs:



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

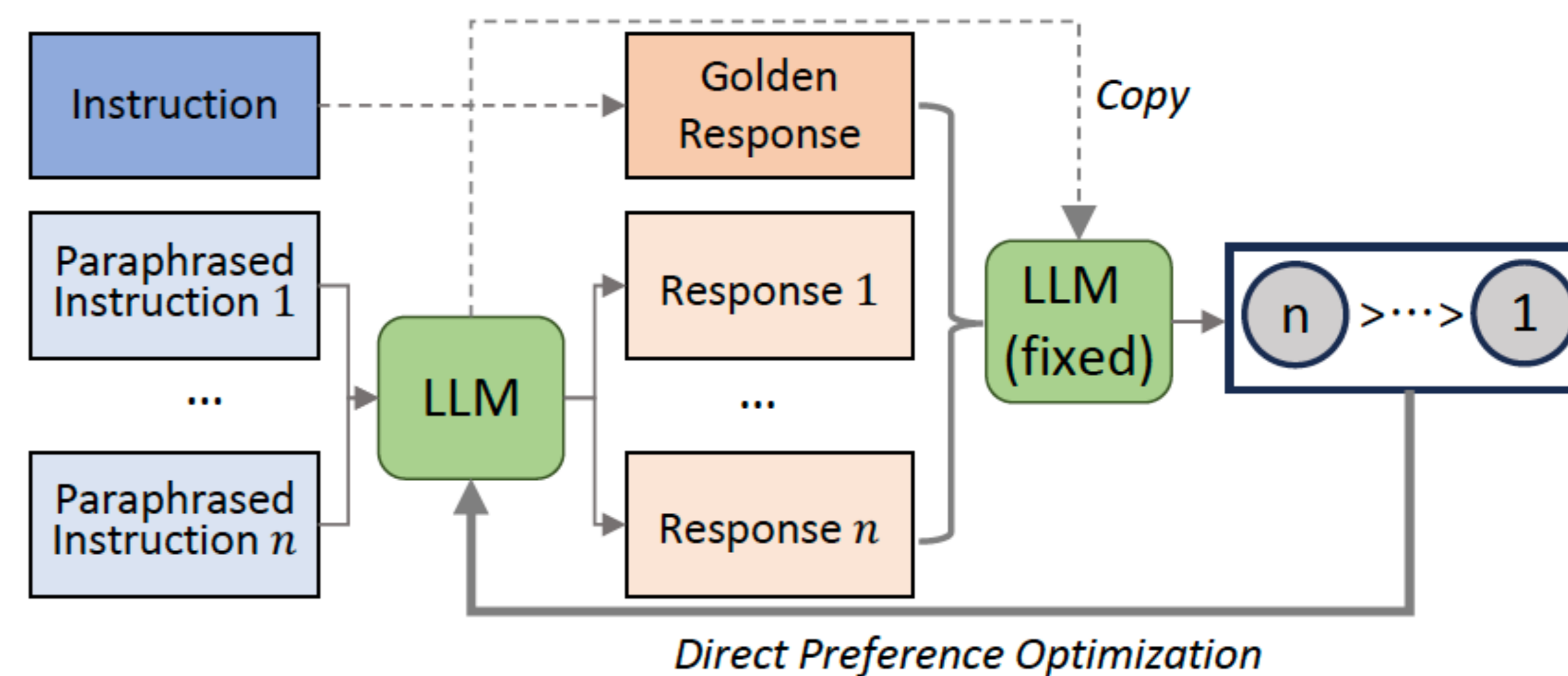
Training Framework

Supervised fine-tuning with instruction augmentation (SFT (IA))



Training Framework

Consistency alignment training with automatic feedback (CAT)



Experiments

- Main results: + SFT > Vanilla; + SFT (IA) > +SFT; + SFT (IA) + CAT > + SFT (IA)

| | 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 |

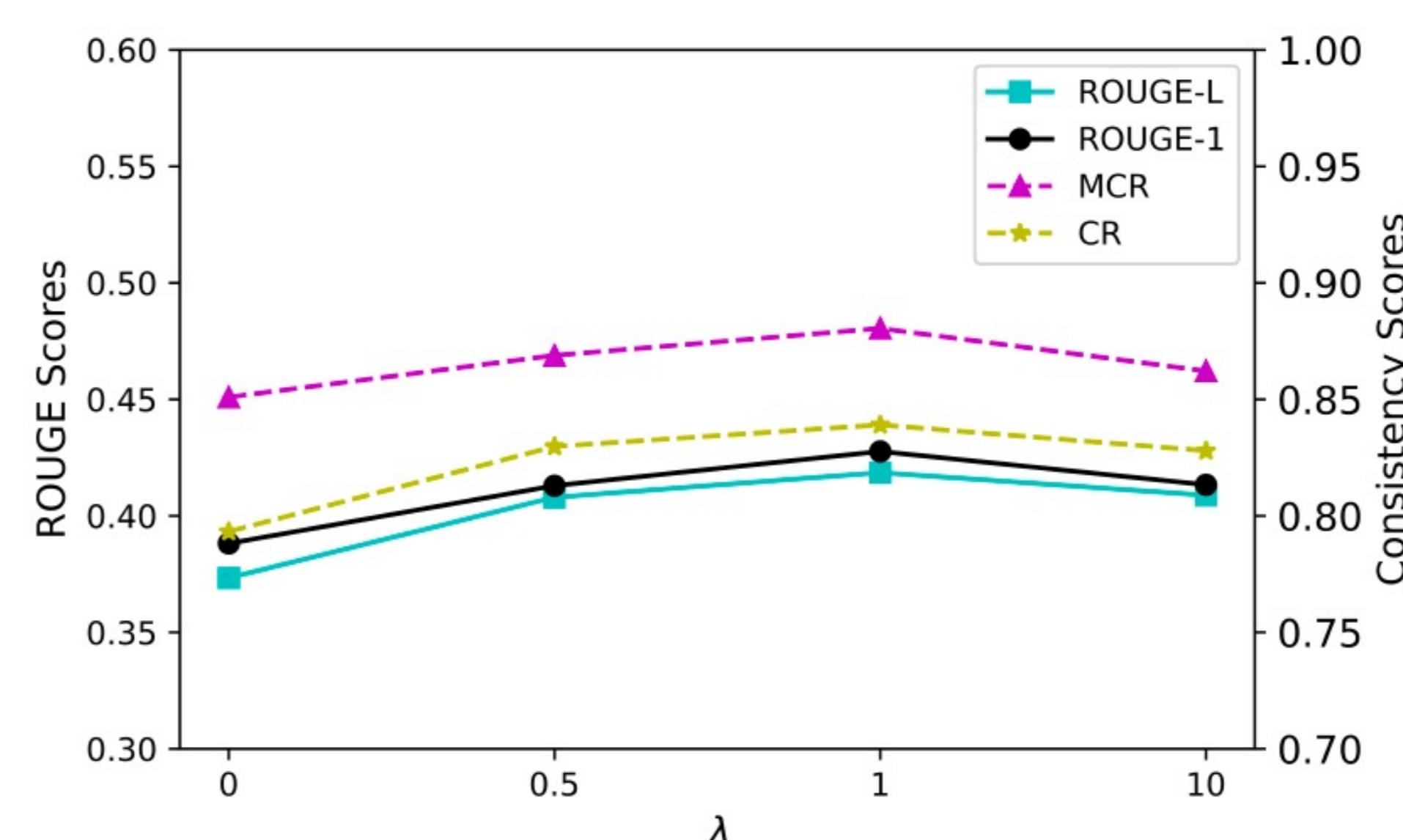
Detailed Analysis

The choice of Rewards

| 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 |

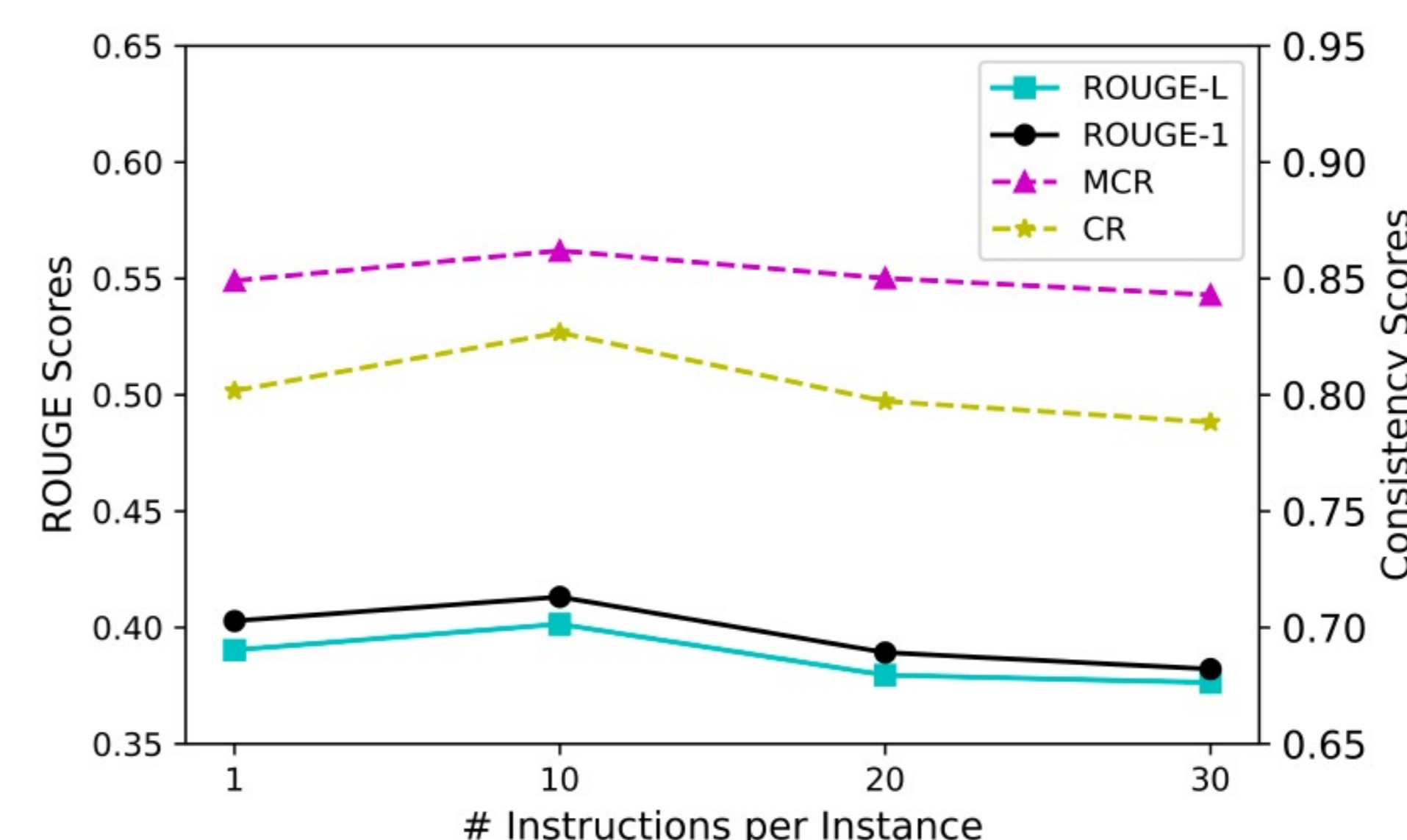
- $r_i^T + r_i^C > r_i^C$;
- A strong LLM is better for rewarding

The choice of λ



- The performance of diff. λ in the loss.
- The necessity of adding the SFT / CAT loss

The performance of diff. number of augmented instructions.



- The necessity of augmenting instructions
- The necessity of sufficient training