Learning LLM-as-a-Judge For Preference Alignment

Ziyi Ye[1], Xiangsheng Li[2], Qiuchi Li[3], Qingyao Ai[1], Yujia Zhou[1], Wei Shen[2], Dong Yan[2], Yiqun Liu[1]

[1]Department of Computer Science and Technology, Tsinghua University

[2]Baichuan AI

[3]University of Copenhagen

2025.3.28

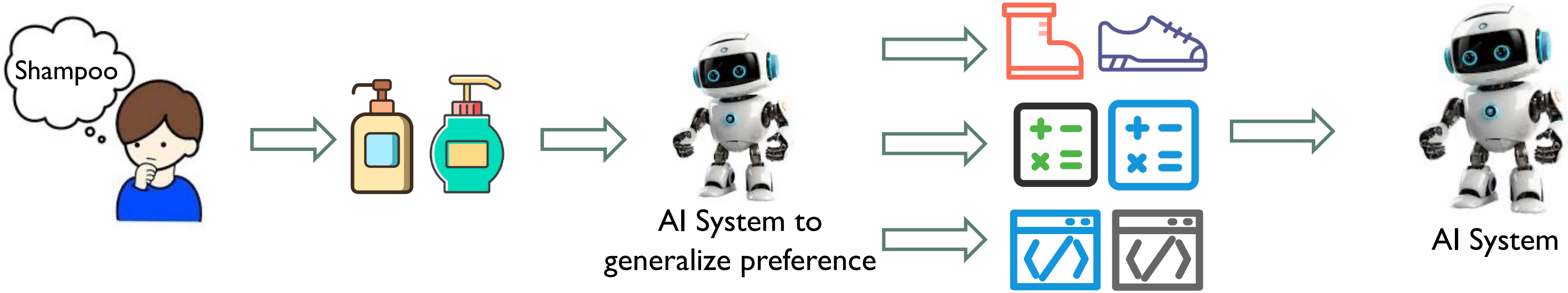
Catalogue

- Background
- Method
- Experimental Results

Background

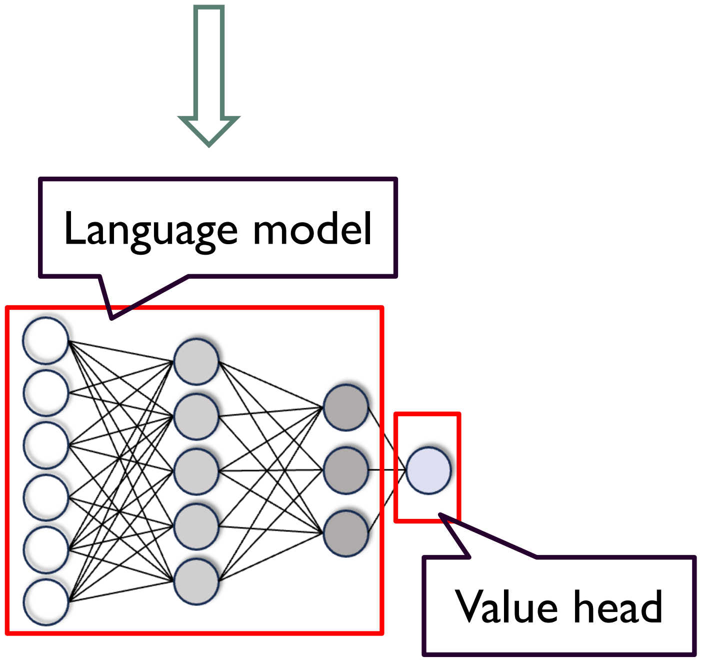
- Learning from preference feedback is a common practice for AI system.

- We need AI agents to generalize human preference for online and infinite tasks.



Background

- A typical solution to generalize human preference in RLHF -> **Scalar Reward model**



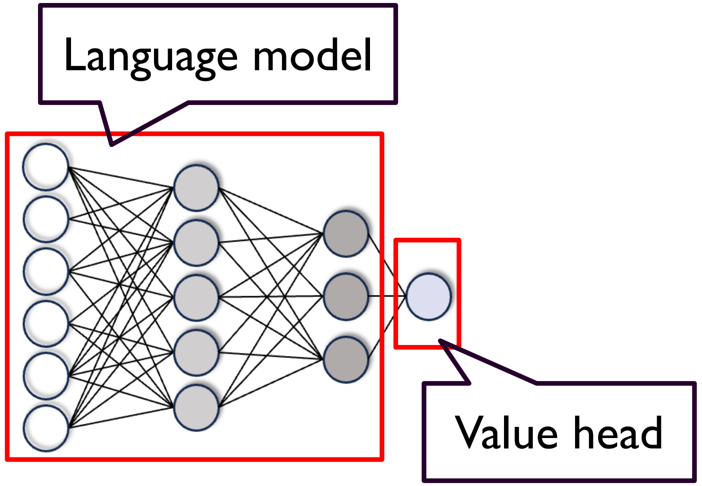
- Challenges for **Scalar Reward model:**

1. Lack interpretability
2. Susceptible to biases

Background

- A typical solution to generalize human preference in RLHF -> **Scalar Reward model**

AI System to generalize preference

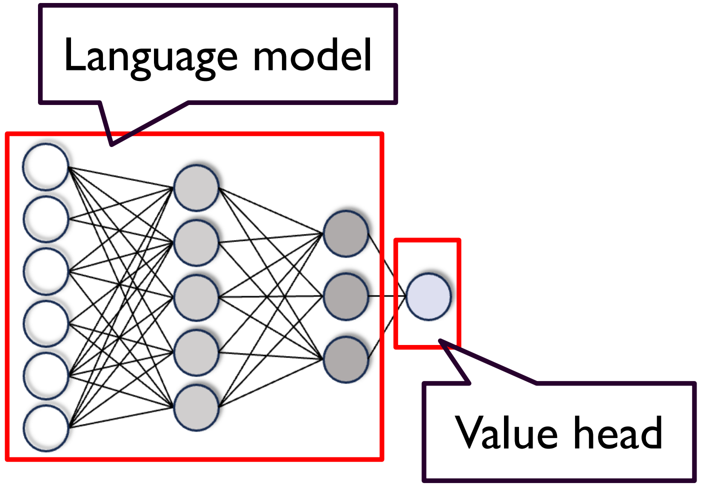


- Challenges for **Scalar Reward model:**

1. Lack interpretability

Background

- A typical solution to generalize human preference in RLHF -> **Scalar Reward model**



AI System to generalize preference

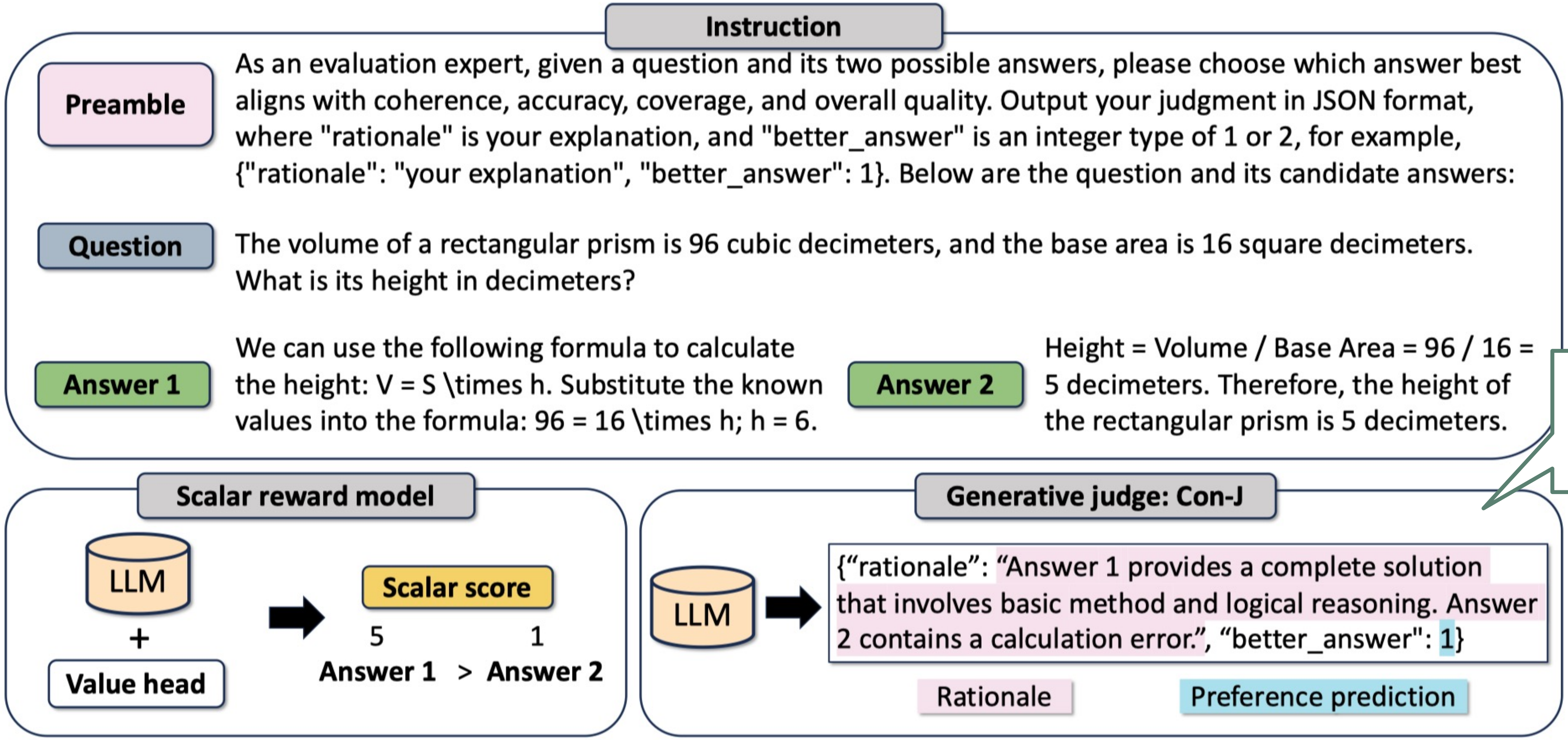
- Challenges for **Scalar Reward model:**

2. Susceptible to biases in datasets

What AI learns: Shampoo with yellow color is better

Motivation

Address **interpretability** and **robustness** in one shot with **Con-J**

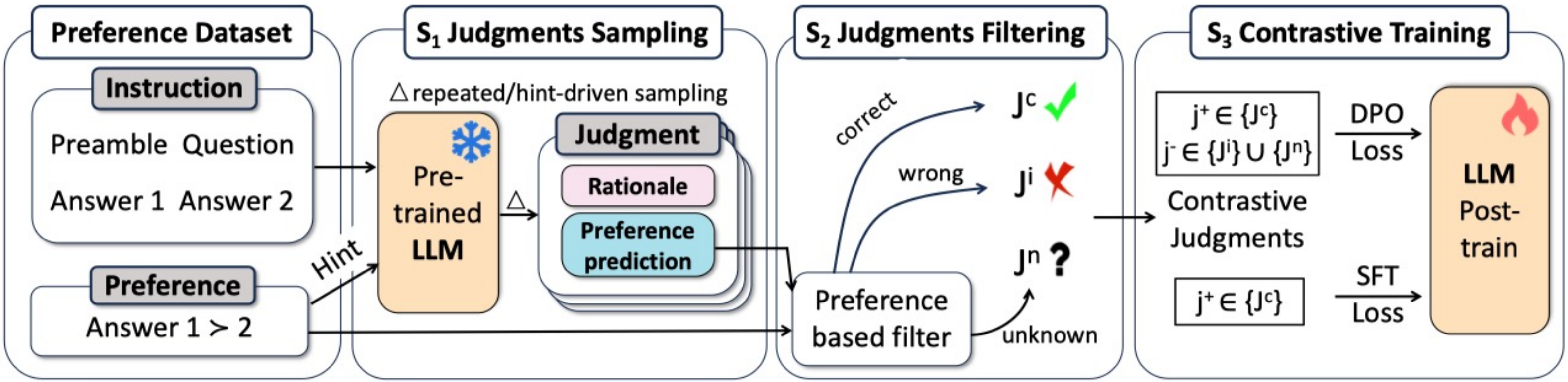


Con-J can generate preference prediction with rationales

Method

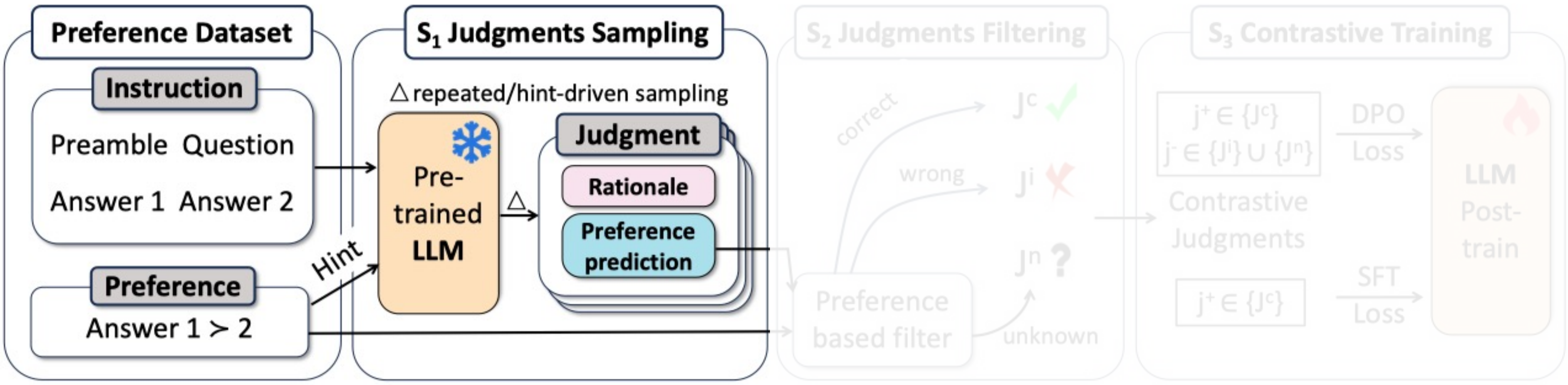
- S_1 Judgments Sampling
- S_2 Judgements Filtering
- S_3 Contrastive Training

Con-J leverages the LLM's pre-existing judgment and bootstraps this ability with human preference



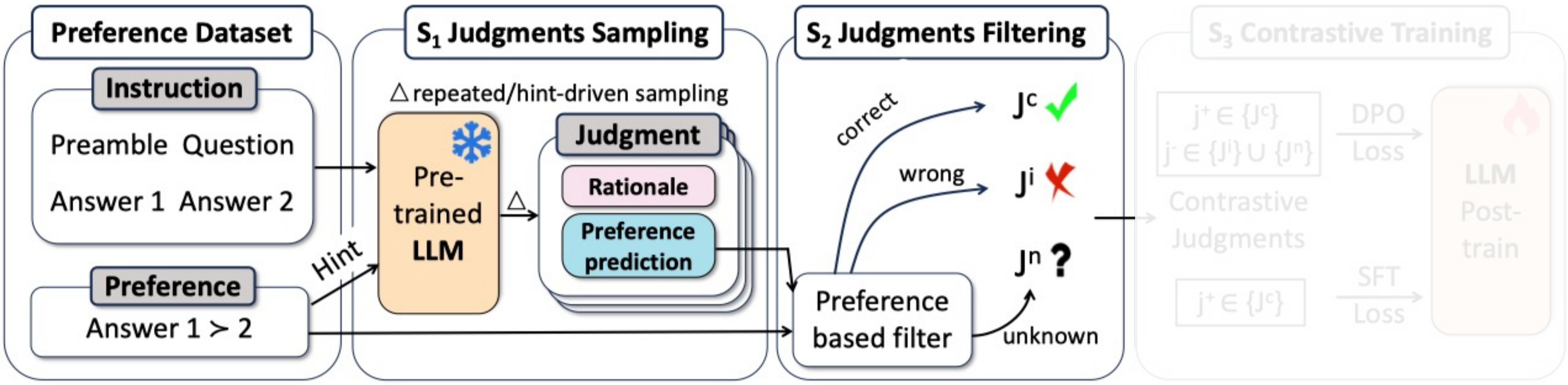
Method

- S_1 Judgments Sampling
  - Repeated sampling
  - Hint-driven sampling -> compel the LLM to generate judgments that prefer specific answers



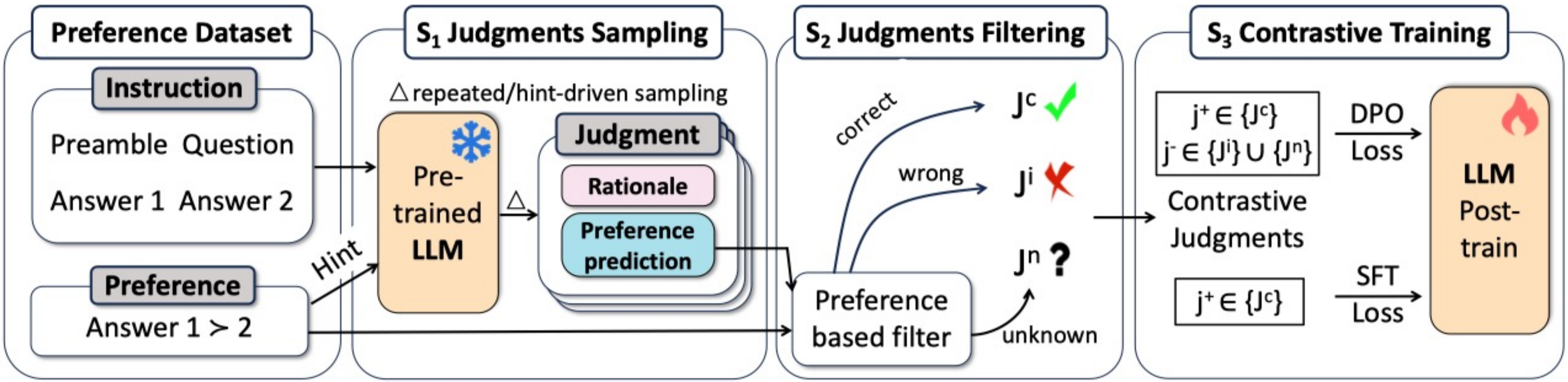
Method

- S_2 Judgements Filtering
  - Positive: the preference prediction corresponding to the keyword "better answer" is correct
  - Negative: the preference prediction is wrong / the format of the answer is incorrect



Method

- S_3 Contrastive Training
  - Train the LLM using DPO loss on positive and negative judgments and SFT loss on positive judgments



Experiment

- Closed source version

Model	Creation	Math	Code
GPT-4o	55.6*	74.8*	68.1*
SM (point-wise)	69.4*	84.8	69.4
SM (pair-wise)	69.2*	84.6	69.6
Con-J	**72.4**	**85.0**	**70.1**

Con-J outperforms scalar reward model (SM) trained on the same corpus and GPT-4o

Model	Creation	Math	Code
Con-J untrained	53.6*	63.4*	61.7*
Con-J w/o Hint	61.3*	77.4*	68.2
Con-J w/o DPO	54.6*	64.2*	63.5*
Con-J	**72.4**	**85.0**	**70.1**

Con-J outperforms its variants w/o DPO and w/o hint-driven sampling

Experiment

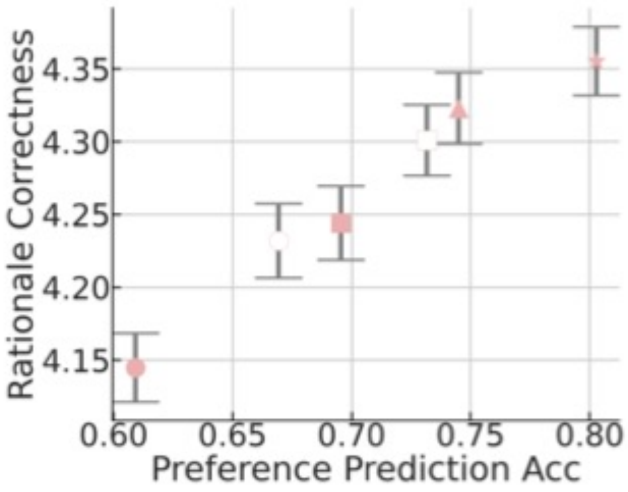
- Open source version

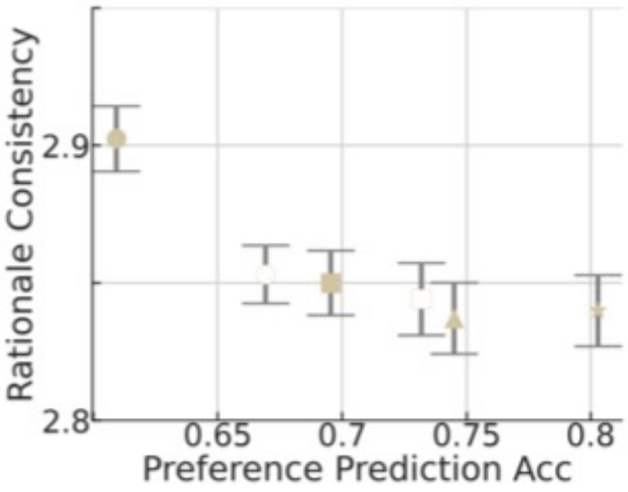
Model	Infinity-Preference	Ultra-Feedback	PKU-SafeRLHF	Reward-Bench Chat	Reward-Bench Chat-H	Reward-Bench Safety	Reward-Bench Reasoning
Llama3.1-8B	59.0	62.9	66.4	80.7	49.8	64.0	68.1
Llama3.1-70B	64.0	71.4	67.6	**97.2**	70.2	82.8	86.0
Qwen2-7B	59.0	64.5	67.2	91.3	44.8	73.6	69.0
Qwen2.5-72B	70.0	66.0	58.7	86.6	61.4	74.5	**90.7**
Auto-J	69.0	63.9	66.9	93.0	40.0	65.5	50.5
Prometheus 2	68.0	63.3	63.0	85.5	49.1	77.1	76.5
GPT-4o	<u>75.0</u>	<u>72.2</u>	**69.6**	<u>95.3</u>	<u>74.3</u>	<u>87.6</u>	86.9
Con-J (ours)	**81.0**	**73.0**	<u>68.4</u>	91.3	**79.6**	**88.0**	<u>87.1</u>

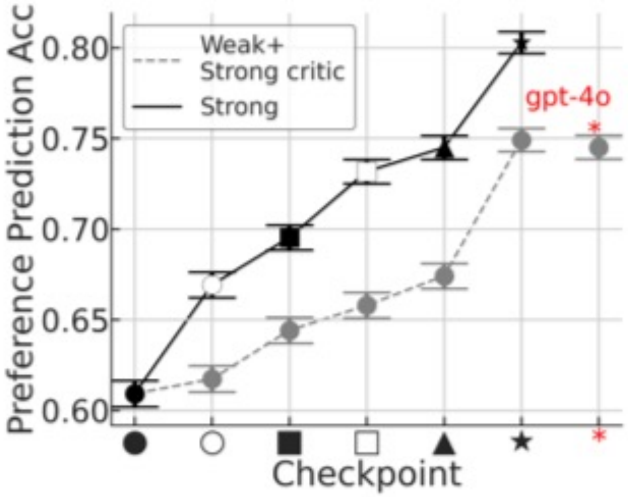
Con-J outperforms or is comparable to state-of-the-art LLM-as-a-Judge

Experiment

- Interpretability of Con-J: is the rationale generated by Con-J useful and reliable?



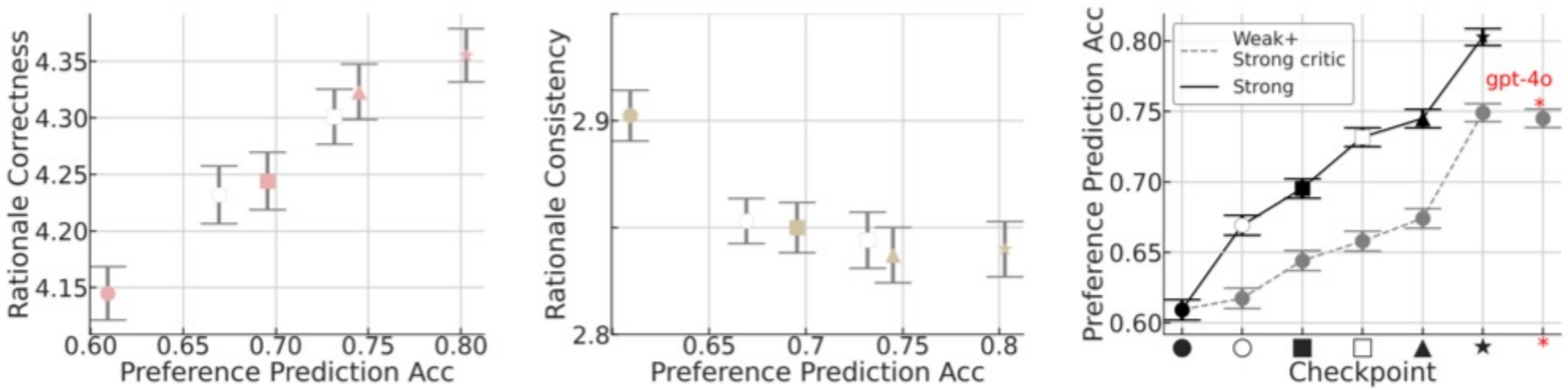




The correctness of the rationales are increasing during Con-J training

Experiment

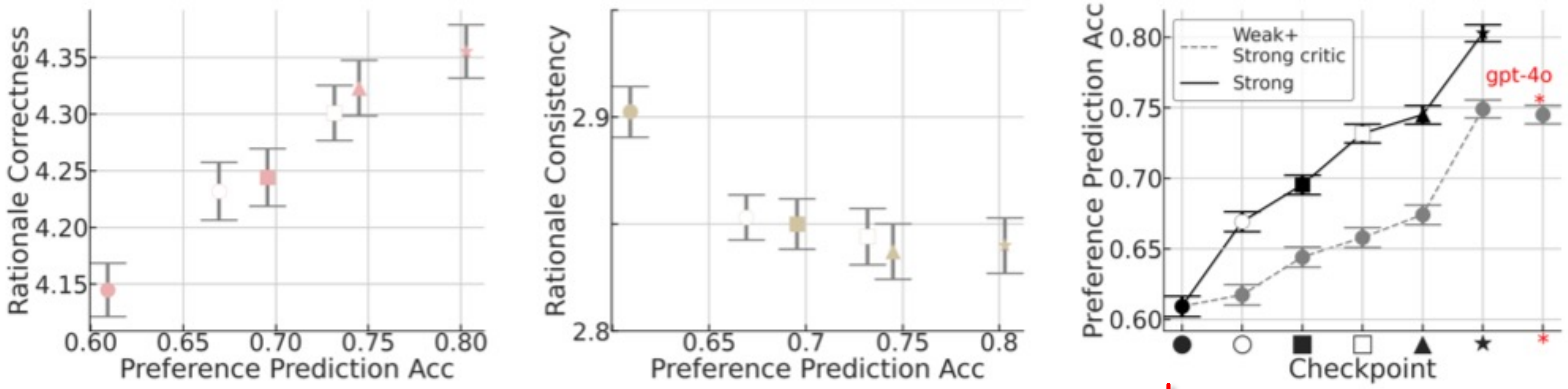
- Interpretability of Con-J: is the rationale generated by Con-J useful and reliable?



The rationales may not be consistent with the final prediction

Experiment

- Interpretability of Con-J: is the rationale generated by Con-J useful and reliable?

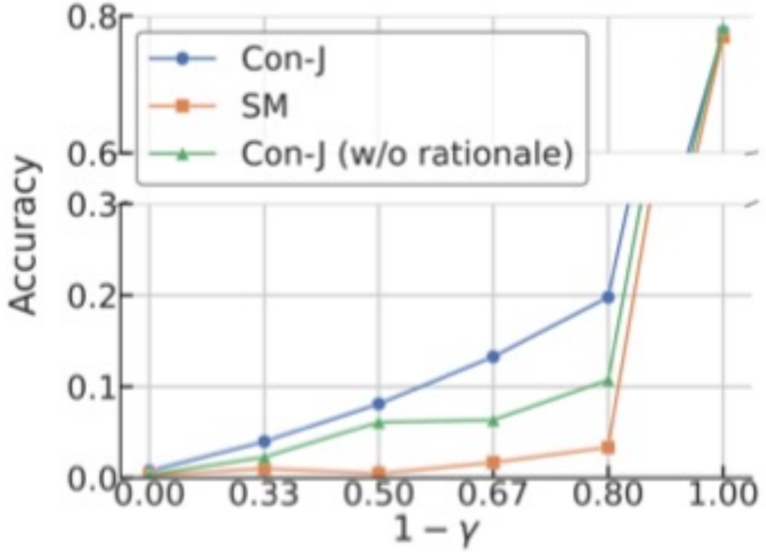


Distilling Con-J's rationale can achieve comparable performance to GPT-4o.

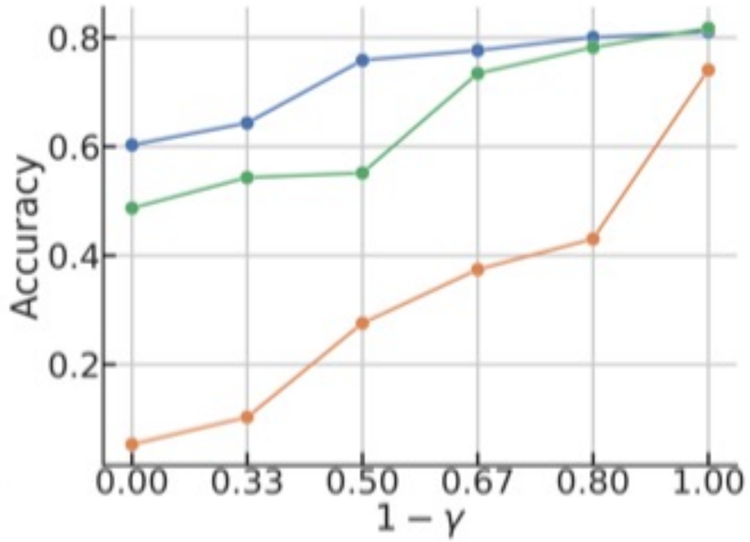
Experiment

- Robustness of Con-J:
  - Why is Con-J more robust than Con-J w/o rationale and scalar model (SM)?

Dataset with format bias



Dataset with verbosity bias

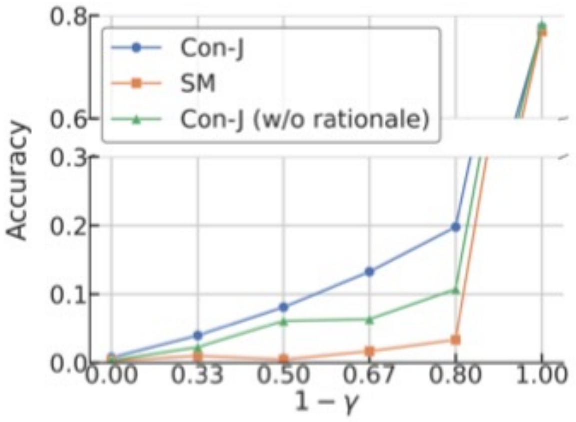


Con-J's trained with rationale is more robust at learning from biased data

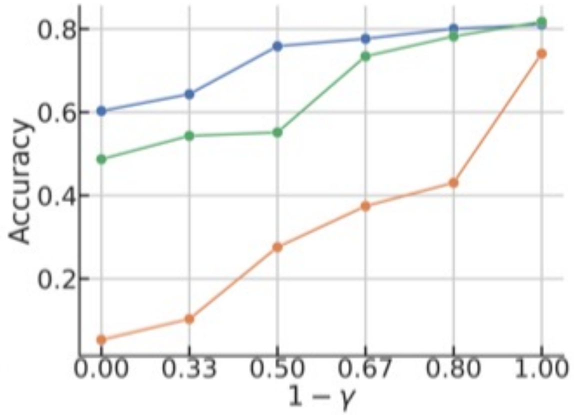
Experiment

- Robustness of Con-J:
  - Why is Con-J more robust than Con-J w/o rationale and scalar model (SM)?

Dataset with format bias



Dataset with verbosity bias



Con-J's trained with rationale is more robust at learning from biased data

Training with rationales bring robustness against bias.

$$P_\theta(j_y \mid p) = \sum_{j_r} P_\theta(j_y \mid j_r, p)\, P_\theta(j_r \mid p)$$

LLM-as-a-Judge resists bias with LLM's prior learning from pretraining. -> a regularization effect

$$\ell(\theta) = \ell_{data}(\theta) + \frac{\lambda}{2}||\theta - \theta_0||^2$$

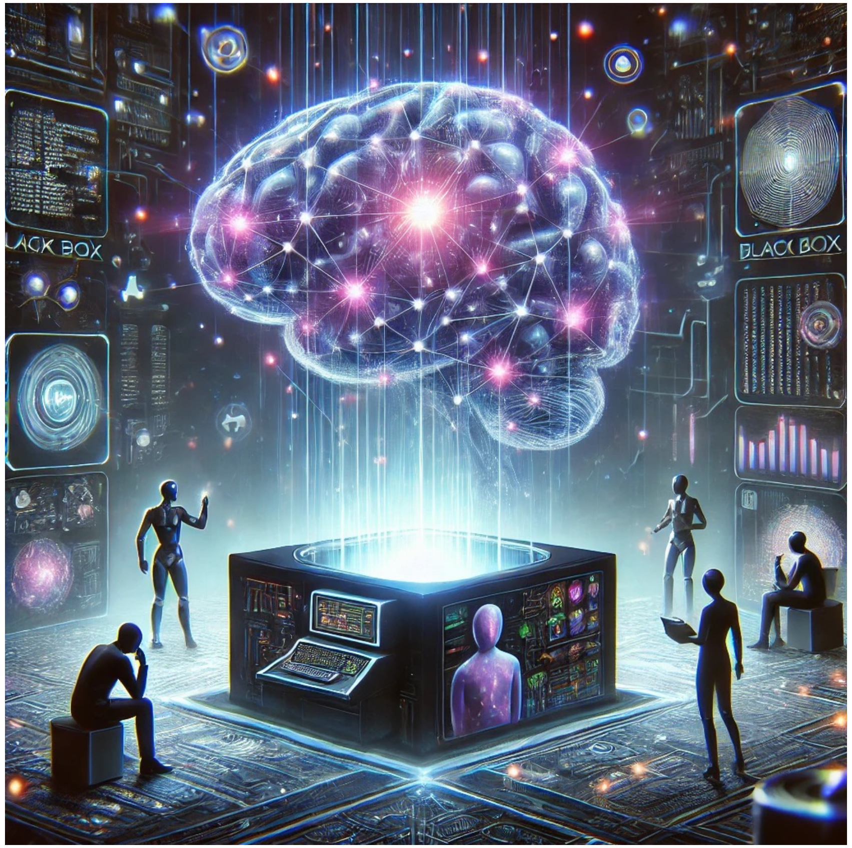
Takeaways

1. **Con-J ->** LLM-as-a-Judge to address the interpretability and vulnerability of scalar reward models

2. **How to train Con-J ->** self-bootstrap, elicit what LLM already knows but supervised by human preference

3. **Result ->** Con-J not only improves its accuracy and robustness in preference prediction but can also generate high-quality rationales

4. **Insight ->** Can we improve AI system in its interpretability and robustness with human preference Signals?



Thank you！ Welcome to our session!

Ziyi Ye[1], Xiangsheng Li[2], Qiuchi Li[3], Qingyao Ai[1], Yujia Zhou[1], Wei Shen[2], Dong Yan[2], Yiqun Liu[1]

[1]Department of Computer Science and Technology, Tsinghua University

[2]Baichuan AI

[3]University of Copenhagen

2025.3.28