SPaR: Self-Play with Tree-Search Refinement to Improve Instruction-Following in Large Language Models

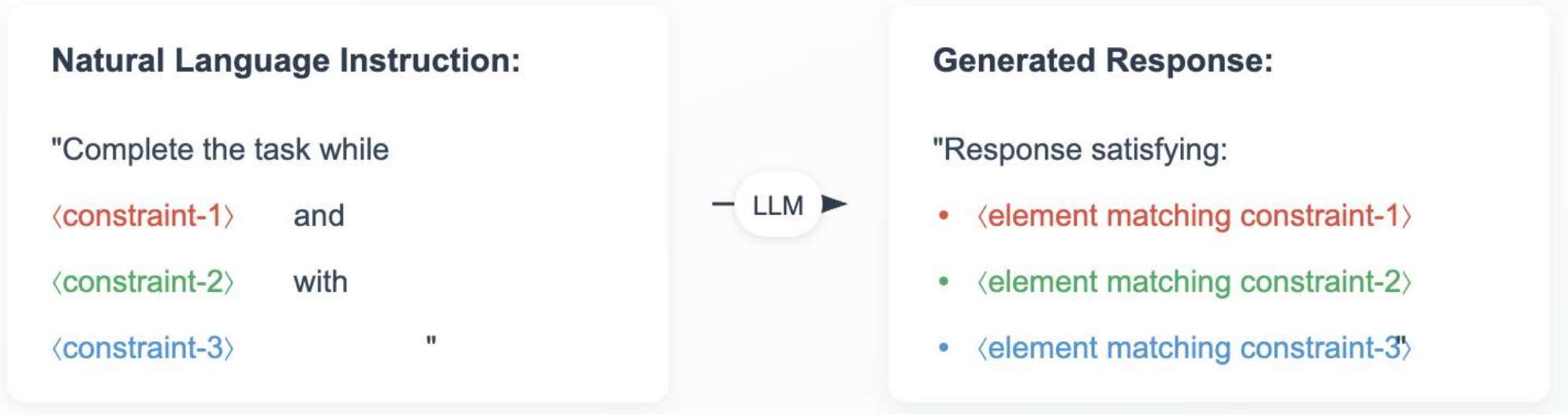
Paper : https://arxiv.org/abs/2412.11605
Data : https://huggingface.co/datasets/CCCCCC/SPaR
Repository : https://github.com/thu-coai/SPaR

Jiale Cheng

清華大學
Tsinghua University

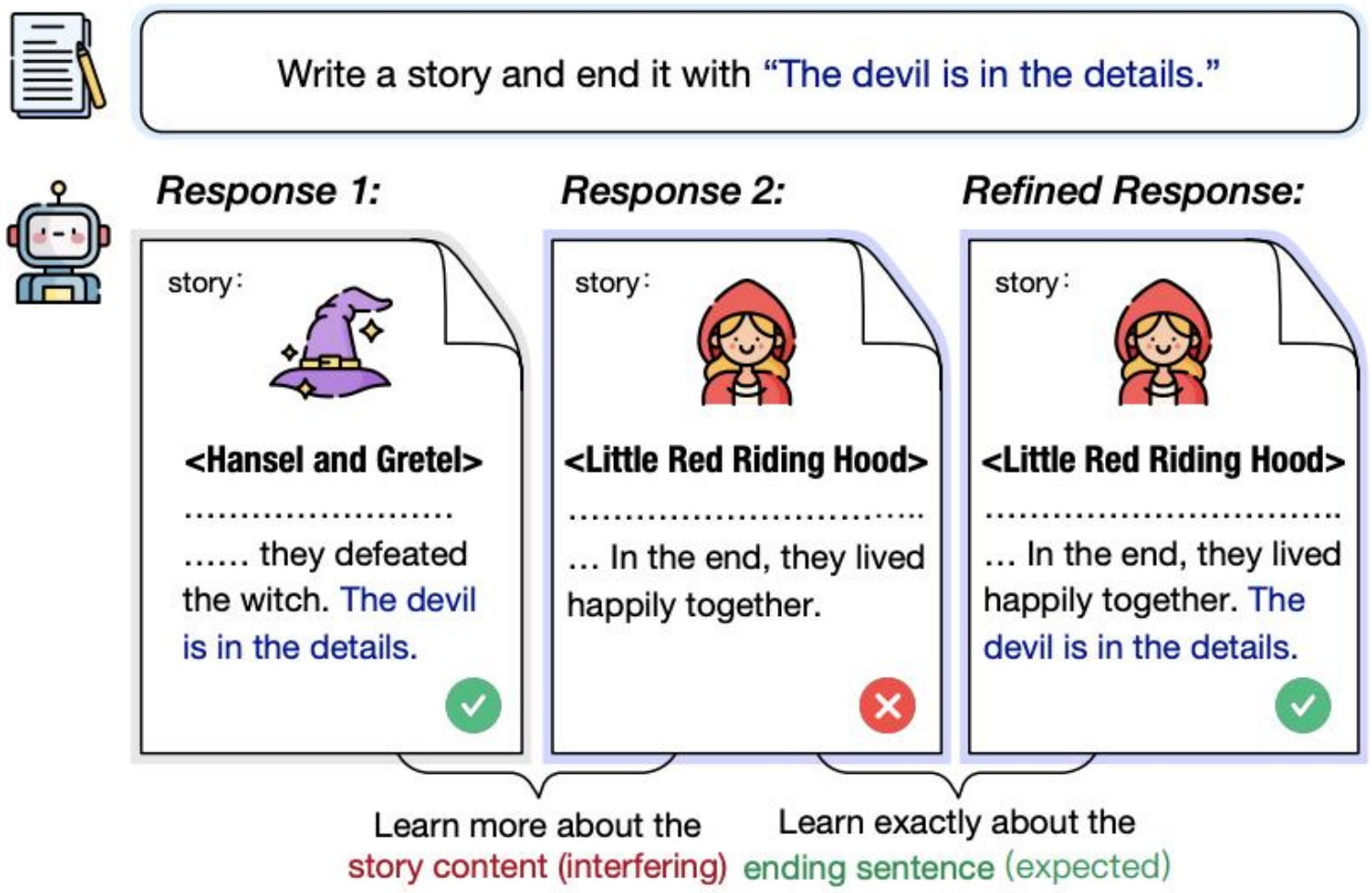
Motivation

- Instruction-Following is the core of LLM alignment, which requires LLMs to recognize subtle requirements and accurately reflect them in responses.



- This ability is well-suited for and often optimized by preference learning.

Motivation



- Challenge:
  - Directly sample multiple independent responses from the model can introduces irrelevant variations (e.g., style or phrasing).
  - This will distract from the core objective of accurately following instructions.

SPaR

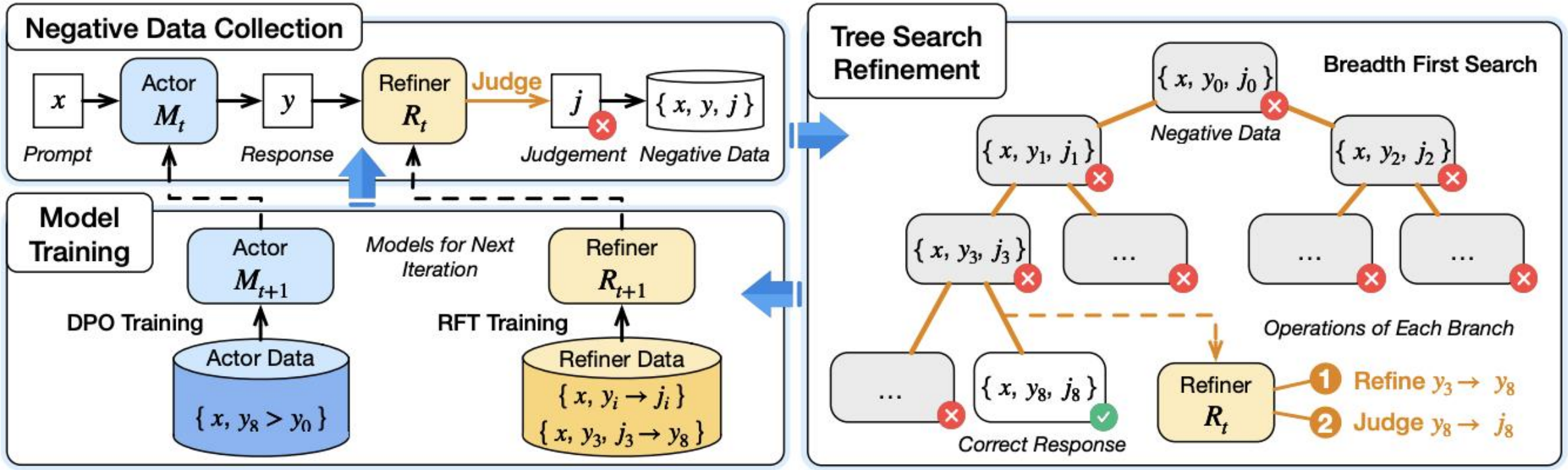
- Our solution: SPaR
  - Self-Play Framework:

    Leverages LLMs playing against themselves to foster continuous self-improvement.
  - Tree-Search Self-Refinement:

    Combines a tree-search strategy with self-refinement to generate valid, comparable preference pairs while reducing irrelevant variations.

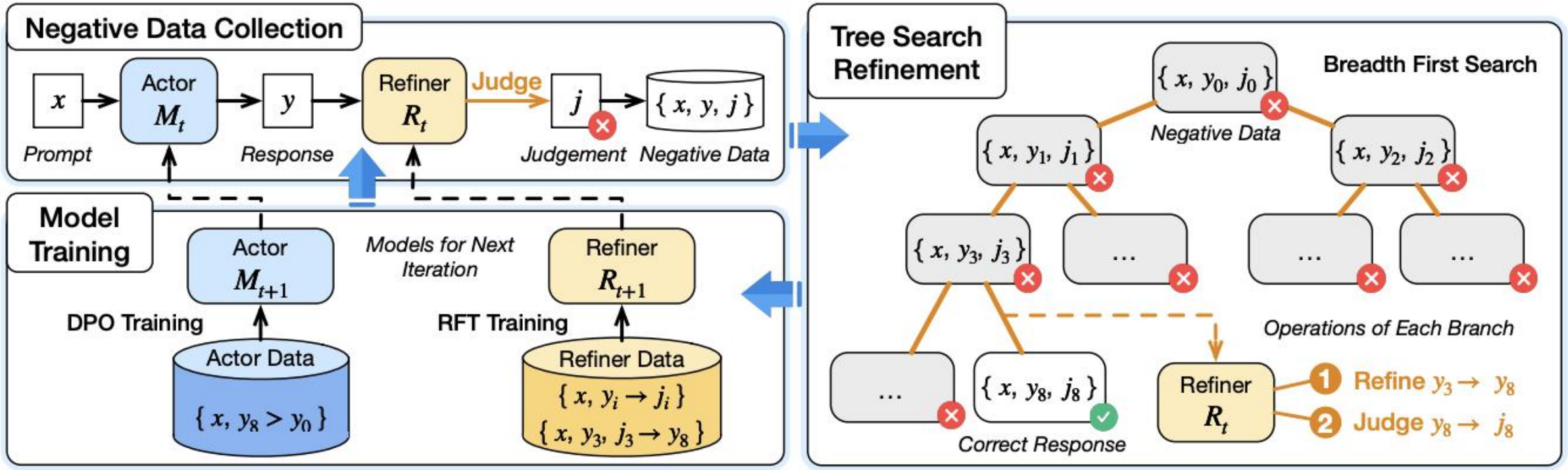
Overall Framework

- Step-1
  - Negative Data Collection: Gather responses that fail to follow instructions accurately.



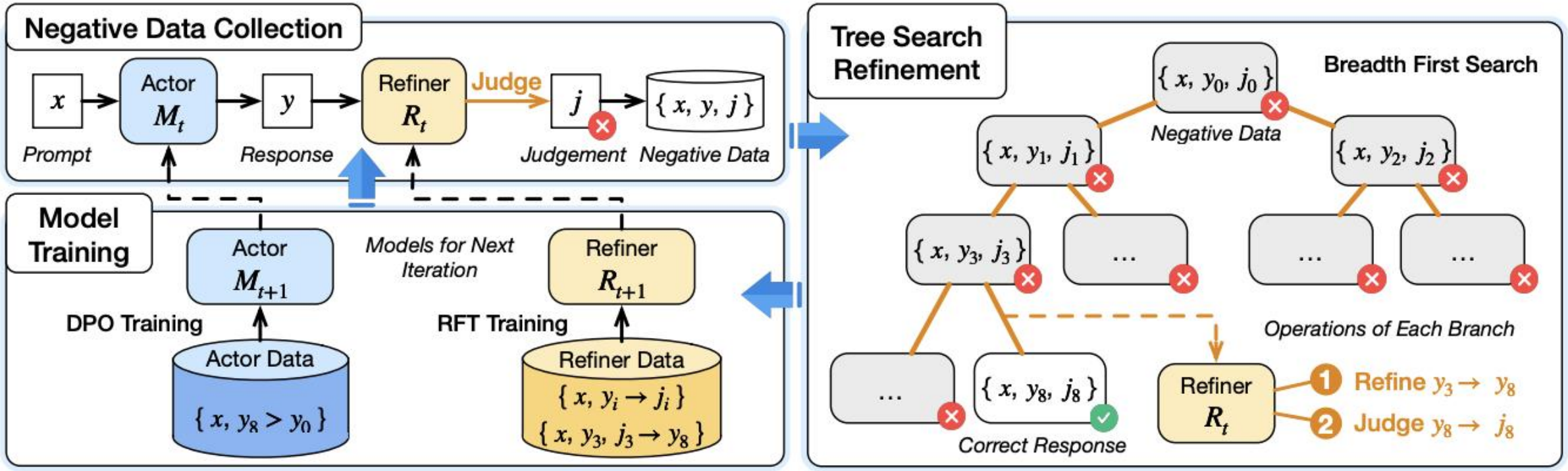
Overall Framework

- Step-2
  - Tree Search Refinement: Apply a tree search to refine these failure cases. Refinement helps exclude interfering factors, while tree search strategy enhances refinement success rate.



Overall Framework

- Step-3
  - Model Training: Train the Actor using DPO and the Refiner using RFT to optimize them and prepare for the next iteration.



Experiment Setup

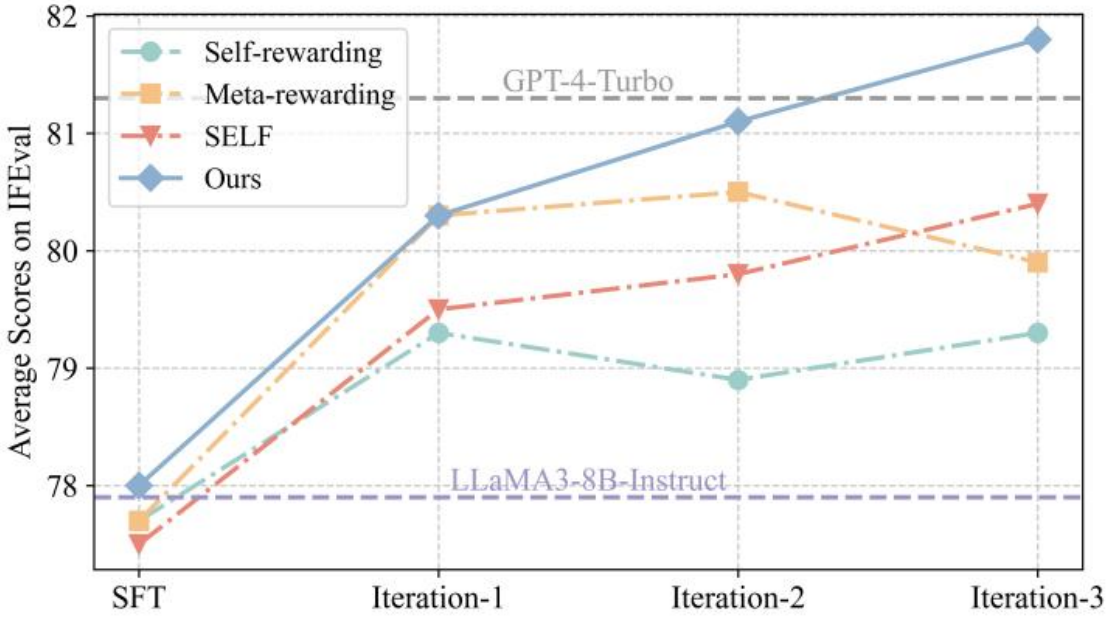
- Test Datasets
  - Instruction-Following Benchmarks: IFEval, FollowBench
  - Instruction-Following Judgment: LLMBar
  - General Ability: GSM8k, TriviaQA, MMLU, HumanEval

- Backbone Models
  - LLaMA-3 series
  - GLM4-9B
  - Mistral-7B

Main Results

- SPaR significantly improves the instruction-following ability of LLMs.

Model	IFEval					FollowBench (SSR)					
	P (L)	I (L)	P (S)	I (S)	Avg.	Lv-1	Lv-2	Lv-3	Lv-4	Lv-5	Avg.
LLaMA3-8B Models											
LLaMA3-8B-Instruct	77.6	84.5	70.6	78.9	77.9	69.4	62.2	63.1	61.9	60.9	63.5
AutoIF-8B†	43.1	56.0	28.8	42.2	42.5	54.6	52.1	50.0	49.0	43.7	49.9
SELF	78.2	84.5	76.0	82.9	80.4	68.3	65.7	65.2	62.2	62.4	64.8
Humpback	72.5	80.2	70.1	78.1	75.2	66.8	66.1	67.2	60.2	62.6	64.6
Self-Rewarding	77.3	84.2	74.1	81.7	79.3	72.8	66.6	66.8	**64.9**	64.1	67.0
Meta-Rewarding	77.8	84.1	75.4	82.3	79.9	73.9	71.9	66.0	62.3	62.6	67.3
SPAR-8B-SFT	75.4	82.5	73.4	80.6	78.0	73.9	67.4	68.1	63.1	61.3	66.8
SPAR-8B-DPO-iter1	78.0	84.7	75.8	82.6	80.3	**75.3**	67.7	67.6	64.7	62.3	67.5
SPAR-8B-DPO-iter2	78.9	85.0	77.1	83.3	81.1	73.9	71.9	69.1	64.0	62.2	68.2
SPAR-8B-DPO-iter3	**79.9**	**85.4**	**78.0**	**83.7**	**81.8**	73.0	**72.3**	**70.0**	64.1	**64.7**	**68.8**
w/ tree search	82.4	87.5	79.5	85.3	83.7	73.9	71.7	70.3	66.8	64.1	69.4
GLM-4-9B Models											
GLM-4-9B-Chat	71.5	79.9	68.0	77.2	74.2	80.8	75.1	67.4	64.3	**65.4**	70.6
SPAR-9B-SFT	71.5	80.5	68.8	78.1	74.7	79.4	70.9	68.2	65.1	63.7	69.5
SPAR-9B-DPO-iter3	**77.3**	**84.1**	**73.6**	**81.4**	**79.1**	**82.7**	**76.7**	**67.9**	**68.3**	64.2	**72.0**
LLaMA3-70B Models											
LLaMA3-70B-Instruct	83.7	88.9	77.1	83.8	83.4	77.1	72.5	69.4	68.7	66.3	70.8
AutoIF-70B†	**85.6**	**90.4**	80.2	86.7	85.7	71.0	67.2	66.2	64.6	63.5	66.5
SPAR-70B-DPO-iter3	**85.6**	90.2	**81.3**	**87.3**	**86.1**	**80.3**	**75.7**	**71.4**	**73.7**	**70.5**	**74.3**



General Performance

- SPaR does not damage general abilities.

Model	GSM8k	TriviaQA	MMLU	HumanEval	Average
			Mistral-7B Models		
Mistral-7B-Instruct	42.9	72.5	57.9	32.9	51.6
SPAR-7B-SFT	56.4	72.8	56.7	44.5	57.6 (+6.0)
SPAR-7B-DPO-iter1	55.6	72.2	55.3	46.3	57.4 (+5.8)
SPAR-7B-DPO-iter2	54.4	72.1	55.8	45.1	56.9 (+5.3)
SPAR-7B-DPO-iter3	58.2	71.6	55.1	46.3	57.8 (+6.2)
			LLaMA3-8B Models		
LLaMA3-8B-Instruct	75.4	75.9	63.6	55.5	67.6
SPAR-8B-SFT	75.6	76.0	64.0	61.6	69.3 (+1.7)
SPAR-8B-DPO-iter1	78.8	75.2	63.8	60.4	69.6 (+2.0)
SPAR-8B-DPO-iter2	77.0	74.9	63.1	60.4	68.9 (+1.3)
SPAR-8B-DPO-iter3	77.7	75.1	63.1	60.9	69.2 (+1.6)
			GLM-4-9B Models		
GLM-4-9B-Chat	80.6	69.7	71.9	74.3	74.1
SPAR-9B-SFT	82.9	69.4	71.8	73.8	74.5 (+0.4)
SPAR-9B-DPO-iter1	82.6	68.8	71.6	75.0	74.5 (+0.4)
SPAR-9B-DPO-iter2	82.8	68.9	71.8	73.8	74.3 (+0.2)
SPAR-9B-DPO-iter3	83.0	69.0	72.1	73.2	74.3 (+0.2)
			LLaMA3-70B Models		
LLaMA3-70B-Instruct	92.2	87.2	80.8	79.3	84.9
SPAR-70B-DPO-iter1	92.5	90.4	81.0	79.3	85.8 (+0.9)
SPAR-70B-DPO-iter2	92.9	89.5	80.4	78.7	85.4 (+0.5)
SPAR-70B-DPO-iter3	93.4	86.7	80.6	79.9	85.2 (+0.3)

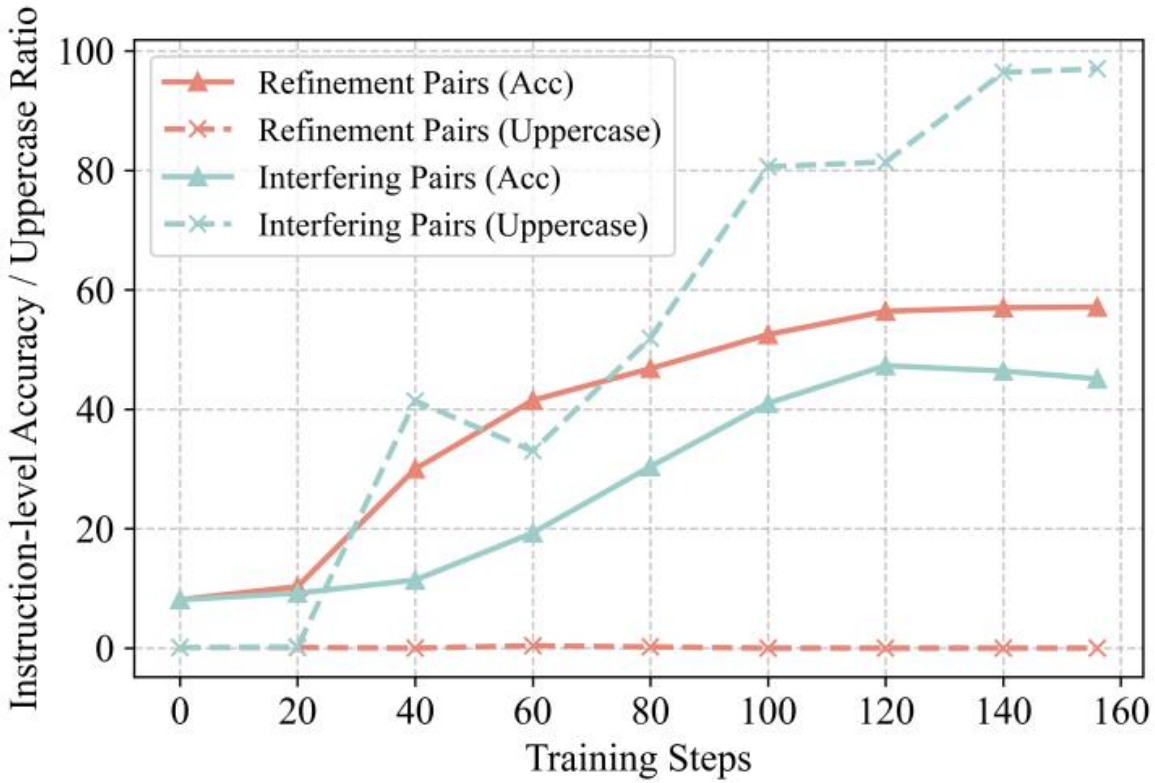
Judgment Evaluation

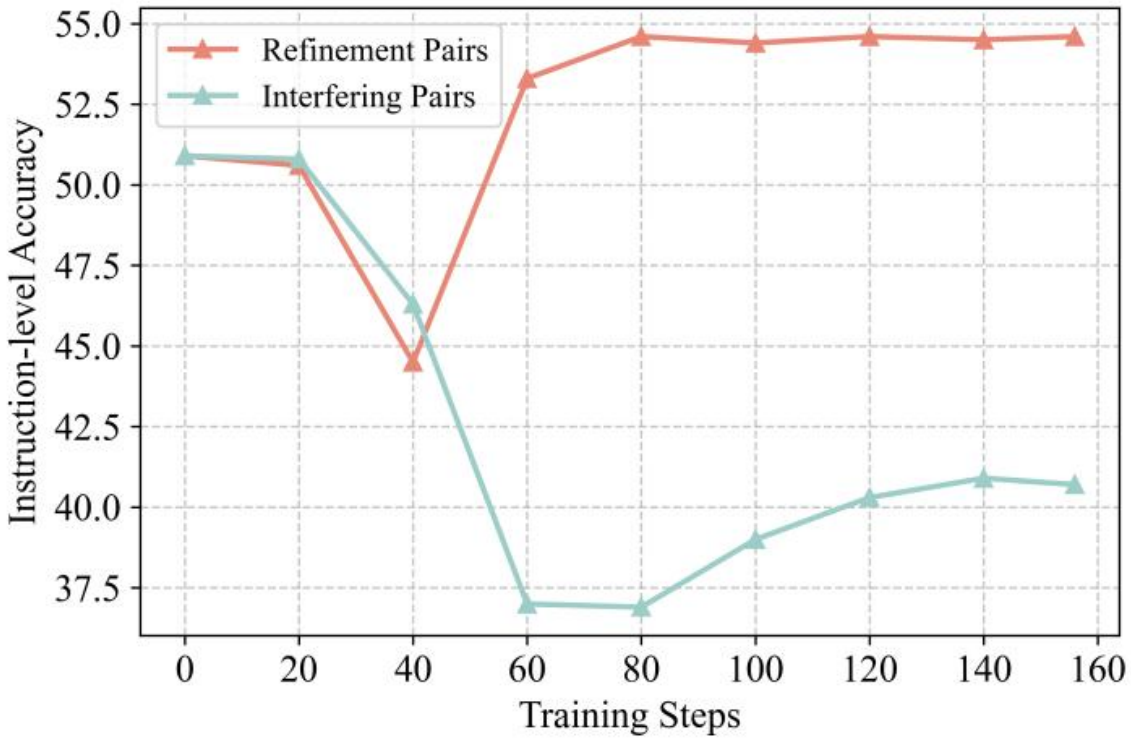
- SPaR iteratively enhances judgment capability.

Model	Natural		Adversarial										Average	
			GPTInst		GPTOut		Manual		Neighbor		Average			
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
GPT-4o-Mini	74.5	70.5	69.2	61.6	60.9	51.4	59.8	51.9	72.8	66.4	65.7	57.8	67.4	60.4
LLaMA3-8B Models														
LLaMA3-8B-Instruct	60.0	51.8	55.4	46.1	47.9	39.5	51.1	36.6	54.5	45.0	52.2	41.8	53.8	43.8
SELF	69.5	61.6	62.0	50.7	64.9	54.8	57.6	41.8	64.6	51.3	62.2	49.6	63.7	52.0
Self-Rewarding	**71.0**	**66.3**	70.1	**66.7**	63.8	59.5	62.0	55.7	67.5	61.7	65.9	60.9	66.9	61.9
Meta-Rewarding	70.5	**66.3**	68.5	64.6	64.9	**60.2**	64.1	58.3	**69.0**	**63.1**	66.6	61.6	67.4	62.5
SPAR-8B-SFT	68.5	60.9	67.9	62.4	59.6	50.0	63.0	54.1	68.3	59.3	64.7	56.5	65.5	57.3
SPAR-8B-RFT-iter1	68.5	63.2	66.8	60.6	63.8	55.3	62.0	53.3	66.8	59.0	64.9	57.1	65.6	58.3
SPAR-8B-RFT-iter2	70.5	64.2	66.8	61.6	**66.0**	60.0	65.2	57.9	**69.0**	62.4	66.8	60.5	67.5	61.2
SPAR-8B-RFT-iter3	70.5	65.9	**70.7**	**66.7**	63.8	57.5	**68.5**	**63.3**	68.3	62.2	**67.8**	**62.4**	**68.3**	**63.1**
GLM-4-9B Models														
GLM-4-9B-Chat	**74.5**	**76.5**	74.5	**75.9**	57.4	**62.3**	53.3	56.6	69.8	**72.0**	63.7	**66.7**	65.9	**68.6**
SPAR-9B-SFT	70.5	65.5	72.8	70.2	**59.6**	55.8	64.1	53.5	71.3	67.2	66.9	61.7	67.7	62.5
SPAR-9B-RFT-iter3	71.0	68.8	**75.5**	74.6	58.5	55.2	**68.5**	**64.2**	**68.7**	65.9	**67.8**	64.9	**68.4**	65.7
LLaMA3-70B Models														
LLaMA3-70B-Instruct	75.0	71.9	73.4	69.6	**69.1**	**66.7**	66.3	**60.8**	69.0	63.4	69.5	65.1	70.6	66.5
SPAR-70B-RFT-iter3	**78.0**	**74.7**	**78.8**	**76.9**	64.9	61.2	**67.4**	59.5	**72.4**	**68.1**	**70.9**	**66.4**	**72.3**	**68.1**

Refinement vs. Random Sampling

- Refinement preference pairs enhance instruction-following capability more effectively.





Ablation & Test-Time Scaling

- Each element is crucial in SPaR.

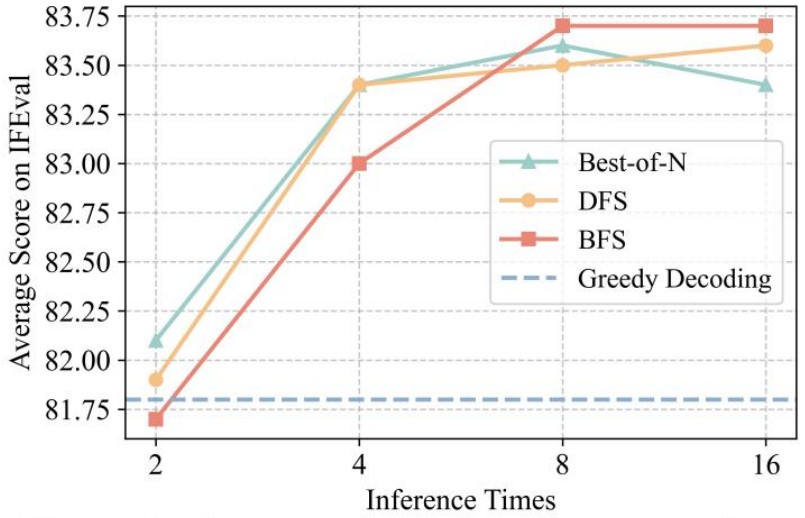
Table 4: Ablation study on the actor.

Model	IFEval		FollowBench (SSR)
	Prompt(S)	Instruction(S)	Avg.
SPAR-8B-DPO-iter3	78.0	83.7	68.8
w/o Tree Search	-2.0	-0.8	-1.7
w/o Iterative Training	-0.9	-0.2	-2.0
w/o Refinement	-2.6	-1.6	-3.1

Table 5: Ablation study on the refiner.

Model	Natural		Adversarial	
	Acc.	F1	Acc.	F1
SPAR-8B-RFT-iter3	70.5	65.9	67.8	62.4
w/o Tree Search	-0.5	-1.2	-4.3	-8.2
w/o Iterative Training	-0.5	-2.5	-1.7	-3.5

- Scaling test-time compute significantly boosts model performance



Conclusion

- Key Takeaways:

- It is important to exclude interfering factors in preference pairs for effective preference learning.
- Iterative self-play enables continuous improvement in instruction-following, judgment, and refinement, creating a scalable, self-reinforcing enhancement loop.
- Test-Time scaling could largely improve instruction-following capabilities.