SPC: Evolving Self-Play Critic via Adversarial Games for LLM Reasoning

Jiaqi Chen

HKU, CS

2025

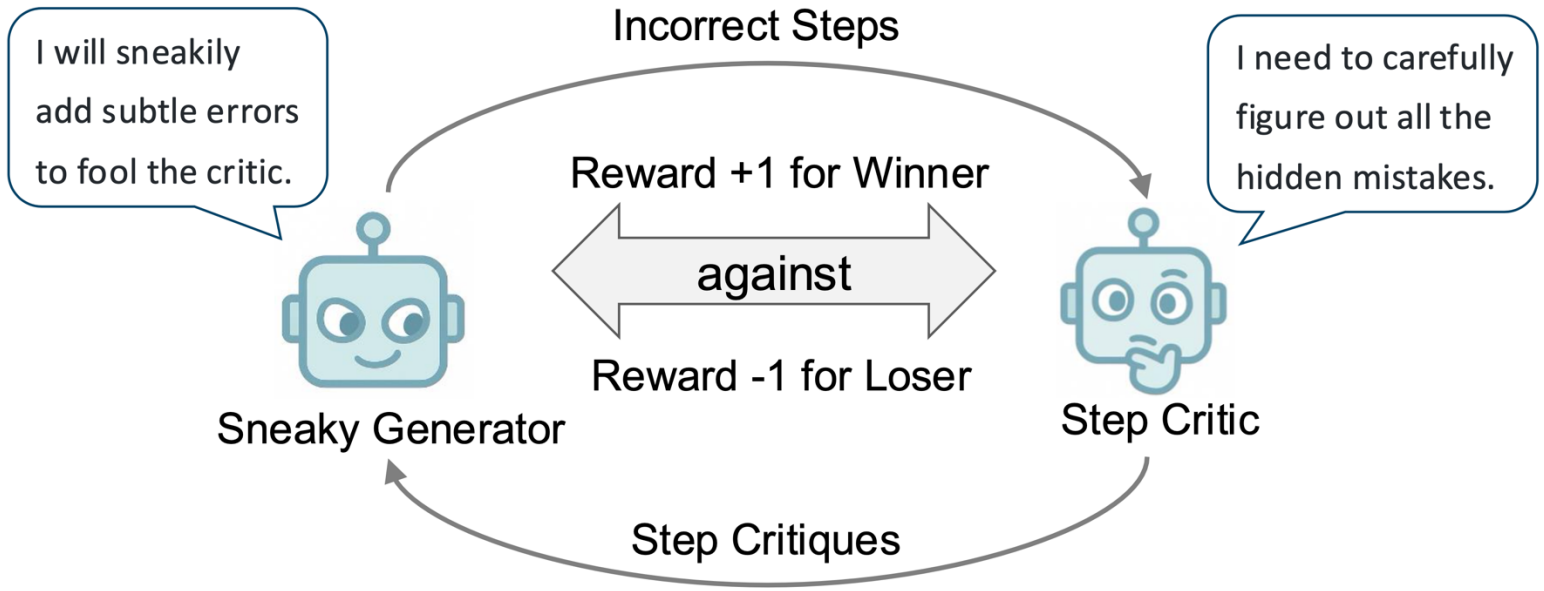
SPC: Evolving Self-Play Critic via Adversarial Games for LLM Reasoning

Complex and Diverse

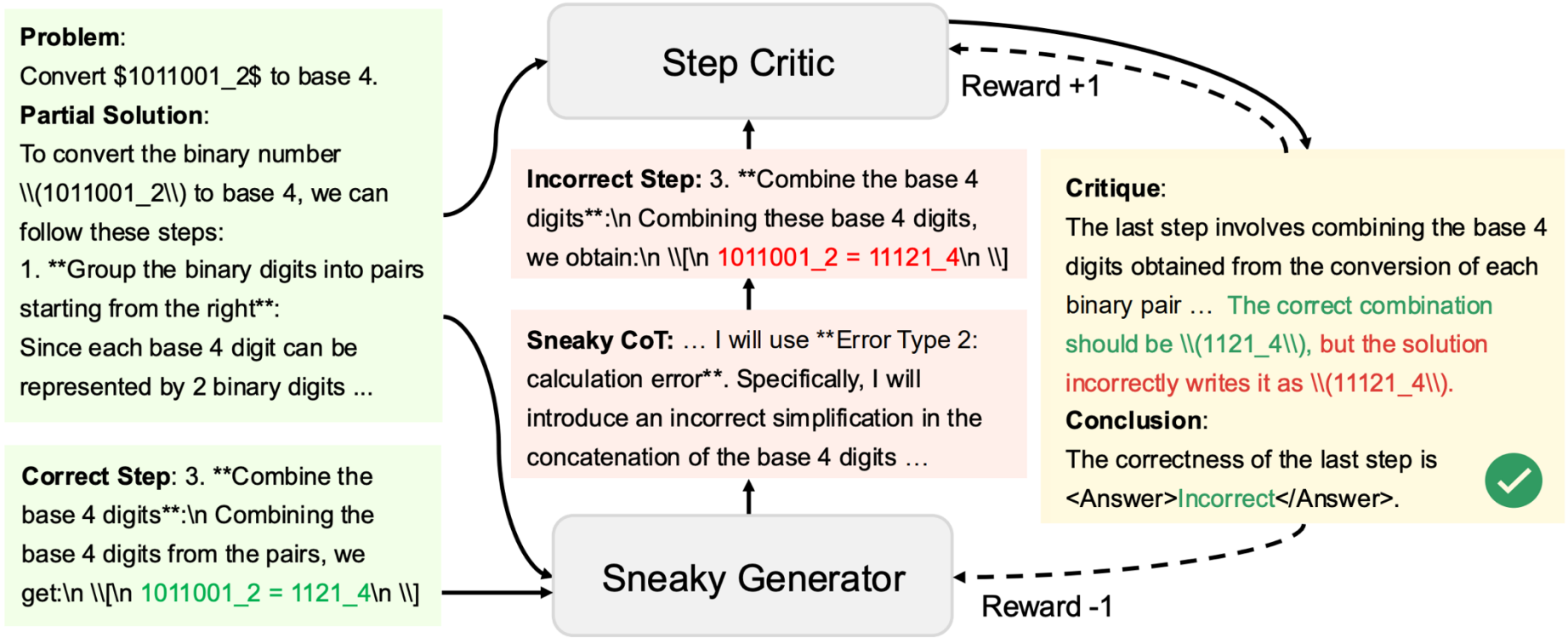
How to evaluate?

Data collection
- Manual labeling
- Automatically generation

SPC: Evolving Self-Play Critic via Adversarial Games for LLM Reasoning



SPC: Evolving Self-Play Critic via Adversarial Games for LLM Reasoning



$$\nabla_\theta \hat{\mathcal{L}}(\theta) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D},\mathbf{y}\sim\pi_{\text{old}}(\mathbf{y}|\mathbf{x})}\left[\frac{\pi_\theta(\mathbf{y}|\mathbf{x})}{\pi_{\text{old}}(\mathbf{y}|\mathbf{x})}\cdot\hat{A}^{\pi_{\text{old}}}(\mathbf{x},\mathbf{y})\cdot\nabla_\theta\log\pi_\theta(\mathbf{y}|\mathbf{x})\right]$$

$$\frac{\pi_\theta(\mathbf{y}\mid\mathbf{x})}{\pi_{\text{old}}(\mathbf{y}\mid\mathbf{x})} = \exp\left(\frac{1}{|\mathbf{y}|}\sum_{t=1}^{|\mathbf{y}|}\log\frac{\pi_\theta(\mathbf{y}_t\mid\mathbf{x},\mathbf{y}_{<t})}{\pi_{\text{old}}(\mathbf{y}_t\mid\mathbf{x},\mathbf{y}_{<t})}\right)$$

$$\hat{A}^{\pi_{\text{old}}} = R(\mathbf{x},\mathbf{y}) - b$$

- 13.2K samples for offline RL
- We consider the average reward of all samples as baseline b

SPC: Evolving Self-Play Critic via Adversarial Games for LLM Reasoning

Models	GSM8K	MATH	Olympiad-Bench	Omni-MATH	Average
Process Reward Models (PRMs)					
Math-Shepherd-PRM-7B [26]	58.0	58.4	68.0	64.1	62.1
Qwen2.5-Math-7B-PRM800K [27]	77.0	72.9	66.9	62.1	69.7
Prompting LLMs as Critic Models					
Llama-3.1-8B-Instruct [10]	59.5	57.7	53.6	53.9	56.2
Llama-3.1-70B-Instruct [10]	67.2	62.8	61.7	61.9	63.4
Qwen2.5-7B-Instruct [12]	64.2	64.0	62.1	60.8	62.8
Qwen2.5-32B-Instruct [12]	76.2	68.1	68.9	63.9	69.3
GPT-4o [6]	75.5	70.5	70.0	64.5	70.1
DeepSeek-R1-Distill-Qwen-7B [21]	79.0	**81.3**	73.4	67.3	75.2
Our Critic Models					
SPC (Round 0)	78.0	74.1	67.8	63.2	70.8
SPC (Round 1)	82.0	80.3	74.8	**70.3**	76.8
SPC (Round 2)	**84.2**	80.8	**76.5**	69.2	**77.7**

Processbench

	PRM800K				DeltaBench			
Models	Average	HarMean	Correct	Error	Average	HarMean	Correct	Error
Process Reward Models (PRMs)								
Math-Shepherd-PRM-7B [26]	50.0	49.5	55.2	44.8	53.3	14.3	7.69	**98.8**
Qwen2.5-Math-7B-PRM800K [27]	73.6	73.6	74.4	72.8	58.5	41.3	**90.1**	26.8
Prompting LLMs as Critic Models								
Llama-3.1-8B-Instruct [10]	51.9	30.5	18.6	85.2	49.1	6.38	3.30	95.0
Llama-3.1-70B-Instruct [10]	54.6	38.9	25.3	83.9	44.6	20.3	11.7	77.5
Qwen2.5-7B-instruct [12]	52.8	37.2	24.1	81.6	48.2	33.8	21.8	74.7
Qwen2.5-32B-instruct [12]	59.0	50.5	36.6	81.4	44.7	33.0	21.8	67.6
GPT-4o [6]	68.5	68.4	70.3	66.6	49.9	48.7	42.0	57.9
DeepSeek-R1-Distill-Qwen-7B [21]	71.4	71.2	67.3	75.5	50.9	50.6	54.9	46.9
Our Critic Models								
SPC (Round 0)	71.0	70.8	67.8	74.2	54.9	53.5	45.9	64.0
SPC (Round 1)	72.8	70.3	59.4	**86.1**	58.8	57.3	68.4	49.3
SPC (Round 2)	**75.8**	**75.8**	**74.8**	76.9	**60.5**	**59.5**	68.2	52.8

PRM800K and DeltaBench

SPC: Evolving Self-Play Critic via Adversarial Games for LLM Reasoning

Table 3: Performance of various methods for assisting different LLMs in math reasoning. By integrating Self-Consistency with our SPC, we achieve the best results across three types of LLMs on MATH500 and AIME2024 datasets.

Solvers	Verifiers	MATH500	AIME2024
Llama-3.1-8B-Instruct [10]	w/o	47.0	4.27
	Self-Consistency [2]	55.6	3.33
	Math-Shepherd [26]	52.4	3.33
	Qwen2.5-Math-7B-PRM800K [27]	54.6	3.33
	Self-Consistency + Math-Shepherd	53.6	6.67
	Self-Consistency + Qwen2.5-Math-7B-PRM800K	60.4	3.33
	SPC (Ours)	54.5	5.63
	Self-Consistency + SPC (Ours)	**62.8**	**6.67**
Qwen2.5-32B-Instruct [12]	w/o	78.0	14.4
	Self-Consistency	82.0	16.7
	Math-Shepherd	78.8	13.3
	Qwen2.5-Math-7B-PRM800K	82.8	16.7
	Self-Consistency + Math-Shepherd	80.8	13.3
	Self-Consistency + Qwen2.5-Math-7B-PRM800K	84.6	16.7
	SPC (Ours)	83.0	17.7
	Self-Consistency + SPC (Ours)	**85.2**	**23.3**
DeepSeek-R1-Distill-Qwen-7B [21]	w/o	87.7	53.8
	Self-Consistency	92.2	70.0
	Math-Shepherd	87.0	53.3
	Qwen2.5-Math-7B-PRM800K	84.2	63.3
	Self-Consistency + Math-Shepherd	89.2	60.0
	Self-Consistency + Qwen2.5-Math-7B-PRM800K	91.8	73.3
	SPC (Ours)	92.3	52.6
	Self-Consistency + SPC (Ours)	**94.0**	**73.3**

THANKS