CasePlay: Self-Play Reinforcement Learning from Case Reports for Medical Reasoning

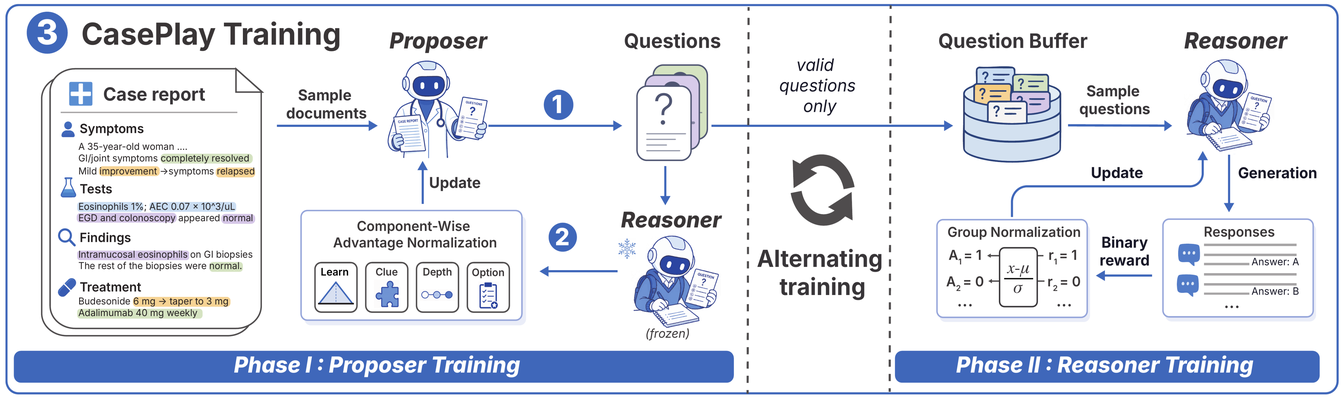
Hao Wang*,1 Zihan Wang*,2 Kai Wu*,3 Jian Yao4 Yiwen Ye5 Yongcan Luo1 Dapeng Wu1 Jiale Chen5 Haihua Yang5

[1]City University of Hong Kong [2]Northeastern University [3]Tongji University [4]The Hong Kong Polytechnic University [5]ByteDance

Contact us: hao.wang@my.cityu.edu.hk

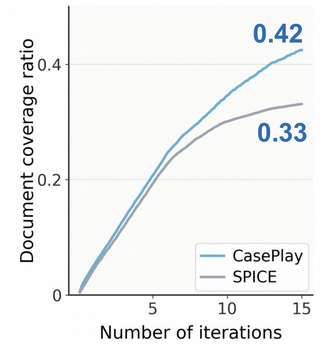
01 Scenario

Can we improve medical reasoning with unlabeled clinical corpora? **Self-play RL is the key.**

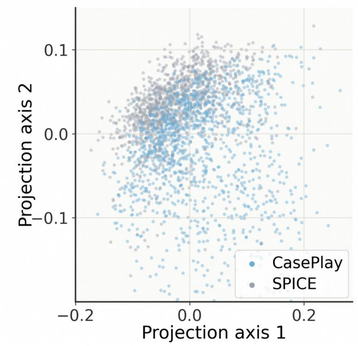


02 Challenges

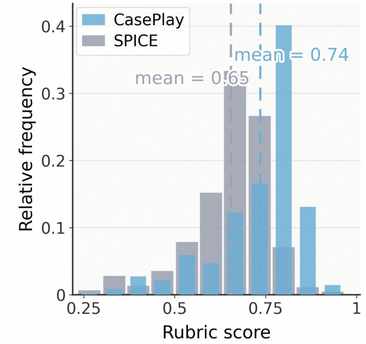
- **Knowledge-point fixation**
- **Difficulty-quality mismatch**



Only 33% doc. coverage after self-play training



Limited diversity of generated questions

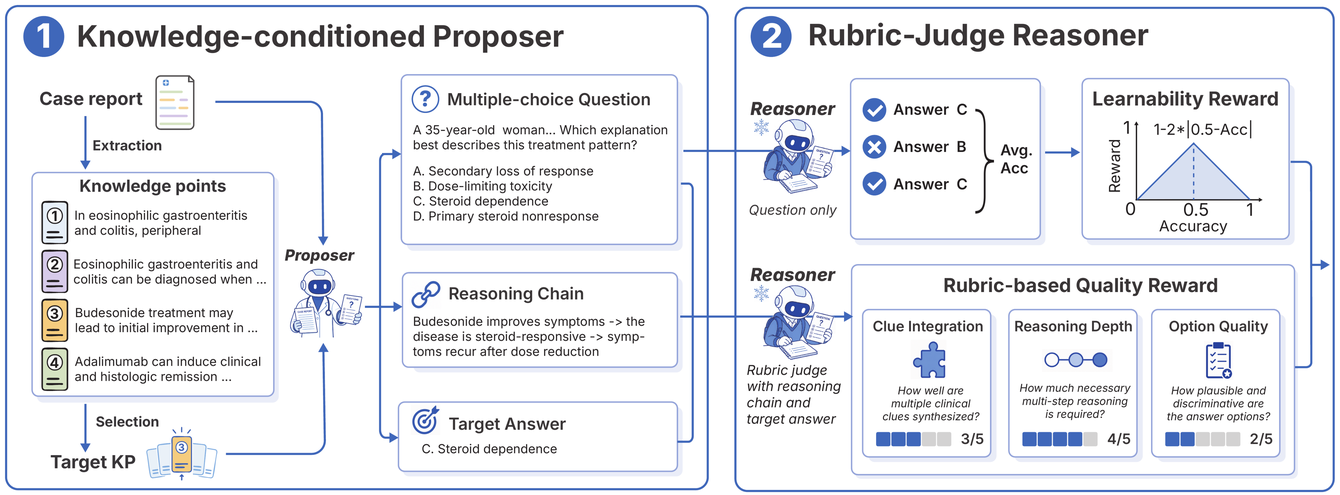


Low-quality questions persist at 50% answer acc

03 Solution

4,281 unlabeled PMC case reports 2020-2025. Shared SFT warm-start for proposer and reasoner.

Iterative GRPO 10 update steps per model per round.



Proposer update Generate high-quality, learnable questions.

$$\hat{A}_i^P = \underbrace{\tilde{r}_i^{\text{learn}}}_{\text{learnability}} + \lambda \underbrace{\left(\tilde{r}_i^{\text{clue}} + \tilde{r}_i^{\text{depth}} + \tilde{r}_i^{\text{option}}\right)}_{\text{rubric feedback from the reasoner}}$$

$r^{\text{learn}} = 1 - 2\,|\text{acc}(q) - 0.5|$ Learnability peaks at 50% accuracy.

Reasoner update Improved by solving questions.

$$\hat{A}_i^R = \frac{r_i - \text{mean}(\{r_{i,g}\}_{g=1}^G)}{\text{std}(\{r_{i,g}\}_{g=1}^G)}$$

$r_i \in \{0, 1\}$ Binary outcome reward

04 Experiments

➢ **How well does CasePlay perform, and do its gains generalize across model families and sizes?**

- **Main Results**

Benchmark	SFT	CasePlay	Gain (pp)
MedXpert	22.7	**25.1**	+2.4
MB-OP4	72.9	**76.9**	+4.0
MB-OP5	68.1	**70.1**	+2.0
MedQA	83.6	**85.4**	+1.8
LiveClin	66.0	**72.3**	+6.3
MMLU-Pro*	75.3	**75.9**	+0.6
GPQA*	53.4	**57.6**	+4.2
Average	**63.1**	**66.2**	+3.1

Medical Reasoning **+2.6 pp** Clinical Case **+6.3 pp** Science QA **+2.4 pp**

- **Self-play baselines**



- **Ablation study**

Variant	Mean (%)	Δ (pp)
No knowledge points	**64.9**	-1.3
No rubric judge	**64.3**	-1.9
Frozen proposer	**63.2**	-3.0

- **Other model scales and families**

Base model	SFT	CasePlay	Δ (pp)
Qwen3-4B-Base	54.4	**58.8**	+4.4
OctoThinker-3B-Hybrid-Base	39.4	**42.1**	+2.7
Qwen3-8B-Base	63.1	**66.2**	+3.1

➢ **Are rubric reliable, and do they guide the proposer training correctly?**

Rubric Dimension	Pearson Corr.
Clue Integration	**0.783**
Reasoning Depth	**0.752**
Option Quality	**0.627**
Average Score	**0.705**

Agreement with DeepSeek-V3.2

