Self-Harmony: Learning to Harmonize Self-Supervision and Self-Play in Test-Time Reinforcement Learning

Ru Wang[1], Wei Huang[2,3], Qi Cao[1], Yusuke Iwasawa[1], Yutaka Matsuo[1], Jiaxian Guo[4]

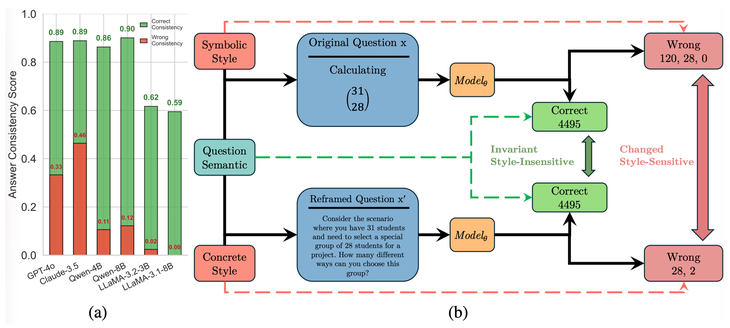
[1] 東京大学 THE UNIVERSITY OF TOKYO [2] RIKEN AIP [3] 大学共同利用機関法人 情報・システム研究機構 統計数理研究所 The Institute of Statistical Mathematics [4] Google Research

ICLR

Background

1. Test-time reinforcement learning (TTRL) offers a label-free paradigm for adapting models using only synthetic signals at inference
2. Its success hinges on constructing reliable learning signals
3. Standard approaches such as majority voting often collapse to spurious yet popular answers.

Observation and Assumption



Observation: Correct answers exhibit significantly higher cross-view consistency than wrong answers across all tested models.
Hypothesis: Because correct reasoning relies on invariant semantics rather than superficial style, the probability of reaching the correct answer C remains stable across paraphrased views:
$P(C \mid View_1) = P(C \mid View_2)$

Theory Analysis

Theorem 3.2 (Harmonic Mean Selector from Invariant Infomax)**.** *Assume the* view-invariance *condition (Assumption 3.1) holds. Suppose moreover that the following conditions are satisfied.*

A1. Non-degeneracy. For every label a, $p_0(a) + p_1(a) < 1$. (This simply excludes the trivial case in where the majority voting can solve.)

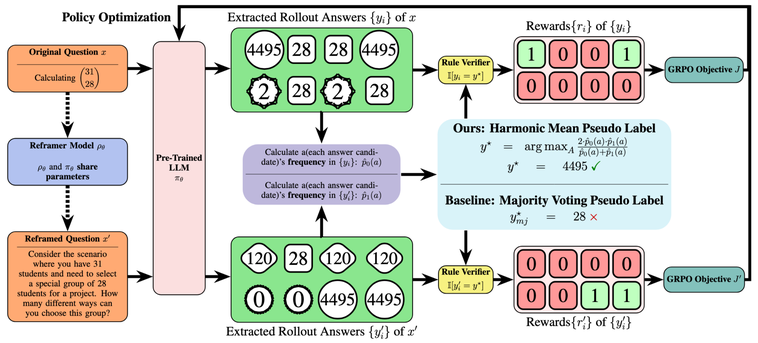
A2. Balanced–Confidence. There exists a constant $\kappa \in (0,1)$ such that for every maximiser $a^\star$ of $J_\lambda(\cdot)$: $\big|p_0(a^\star) - p_1(a^\star)\big| \le \kappa\big[p_0(a^\star) + p_1(a^\star)\big]$.

A3. Uniform View Prior. The view variable $X \in \{x, x'\}$ is assumed to be sampled uniformly, i.e., $p(X = x) = p(X = x') = \frac{1}{2}$. This assumption ensures that the penalty term $I(Z_a; X)$ reflects dependence on the view itself rather than bias in the sampling distribution.

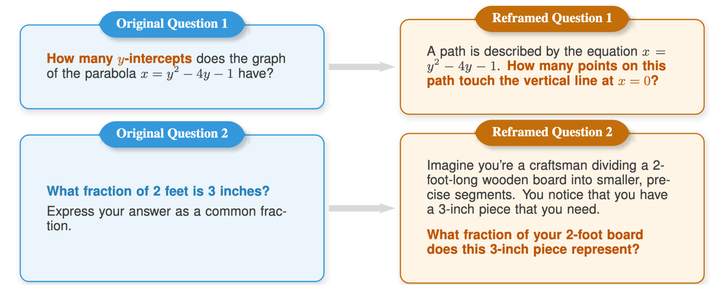
Then, for the penalty weight $\lambda = 2$, the pseudo label that maximises the second-order approximation of the view-invariant Infomax *objective $J_2(a) = I(Z_a; A) - 2\, I(Z_a; X)$ is obtained by the harmonic mean of the two view-probabilities:*

$$y^\star = \arg\max_a \frac{2\, p_0(a)\, p_1(a)}{p_0(a) + p_1(a)} \in \arg\max_a J_2(a)$$

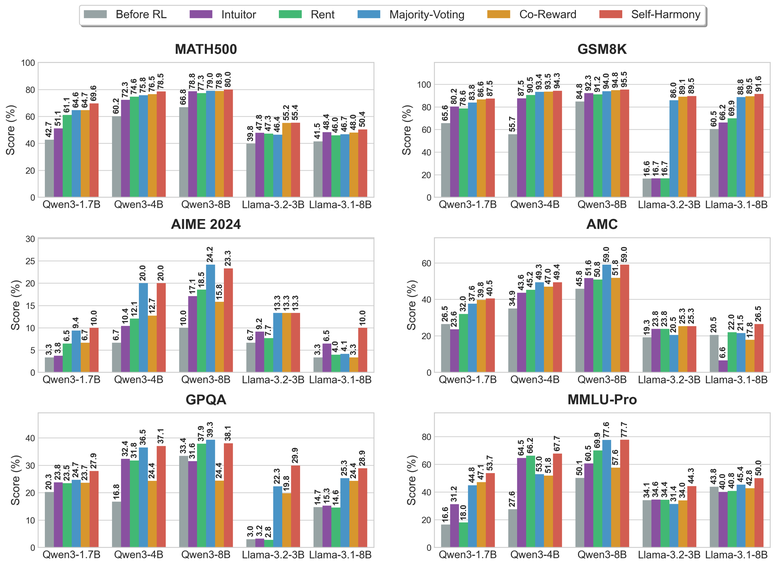
Self-Harmony



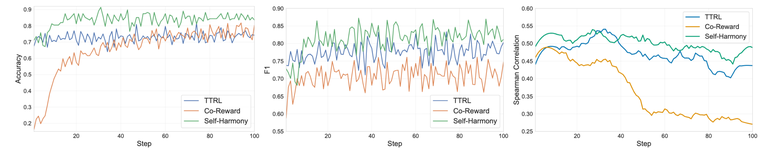
Original and Reframed Questions



Main Result



Pseudo-Label Analysis

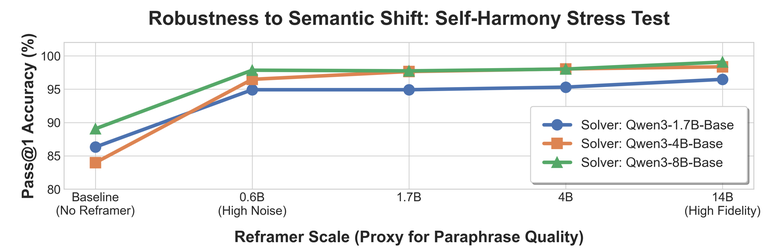


(a) Pseudo-Label Accuracy (b) Pseudo-Label F1 (c) Pseudo-Label Spearman(Sm.)

Harmonic Mean Interpretation

$$\frac{2p_0(a)p_1(a)}{p_0(a)+p_1(a)} = \frac{1}{2}\left[\underbrace{(p_0(a)+p_1(a))}_{\text{Self Consistency}} - \underbrace{\frac{(p_0(a)-p_1(a))^2}{p_0(a)+p_1(a)}}_{\text{View Invariance}}\right]$$

Semantic Shift Pressure Test



Ablation Study

Math500	Self-Harmony (Full)	Reframer		Pseudo-label Selection	
		w/o Format Reward	w/o Diversity Reward	Cross Selection	Majority Voting
Qwen3-4b-Base	**78.50**	78.40	78.20	76.50	77.30
Qwen3-8b-Base	**79.80**	77.46	78.90	78.40	79.00

Take Home

1. Theory proves the Harmonic Mean beats Majority Voting under our assumptions, we encourage using it as a robust drop-in replacement for *any* pseudo-labeling pipeline.
2. A single-model Solver-Reframer generates diverse views to expose fragile reasoning, with extended theories covering scenarios that violate standard view assumptions.
3. Self-Harmony achieves SOTA unsupervised adaptation across 5 models and 6 tasks

Paper **Code**