

Test-Time Training on Binary Sub-Problems

Andrew Sung¹ Darrow Hartman¹ Leo Li¹

¹Department of Computer Science, Stanford University (Equal Contribution)



Problem: Binary Rewards Break Test-Time RL

- Test-time training (TTT) lets LLMs keep learning at inference on a single problem
- TTT-Discover [1] achieved SoTA using continuous rewards
- Math, code, and verification have only **binary feedback**: right or wrong

Problem too hard $\Rightarrow$ all rollouts score $R = 0$
All GRPO advantages = 0 $\Rightarrow$ **no learning signal**

Standard GRPO on the target directly **cannot work**: at 0.67% solve rate, virtually every rollout group scores 0—no reward variance, no gradient. **Direct GRPO produces no improvement**.

Our idea: Break the hard problem into **learnable subproblems**, use model consensus as ground truth, train with GRPO on those instead.

GRPO optimization [2]: for each subproblem q , sample G responses from π_θ and assign rewards $r_i \in \{0, 1\}$ via exact match. Compute group-normalized advantages and update via clipped surrogate:

$$\hat{A}_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r}) + \epsilon}, \quad \rho_i = \frac{\pi_\theta(o_i | q)}{\pi_{\text{ref}}(o_i | q)}$$
$$\mathcal{J} = \frac{1}{G} \sum_{i=1}^G \min\left(\rho_i \hat{A}_i, \text{clip}(\rho_i, 1 \pm \epsilon) \hat{A}_i\right) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})$$

When **all** $r_i=0$: $\text{std}(\mathbf{r})=0$, $\hat{A}_i=0 \forall i \Rightarrow \nabla_\theta \mathcal{J} = \mathbf{0}$.

Method: Pipeline



~ 1 in 7 candidates pass filtering; ~ 586 discarded outside the 60–80% window.

Key Insight: Why 60–80% Agreement?

Self-consistency filtering serves a **dual purpose**: difficulty calibration *and* pseudo-label generation. For binary rewards, let p denote the model's solve rate on a subproblem:

- Reward variance**: $\text{Var}(r) = p(1-p)$, maximized at $p=0.5$. GRPO gradient magnitude scales with $\text{std}(r)$ [3].
- Label trust**: The modal answer becomes the pseudo-label. Below 60% agreement, the majority is too thin to trust.
- Our window balances both**: we shift above the theoretical optimum of 50% to ensure pseudo-label reliability.

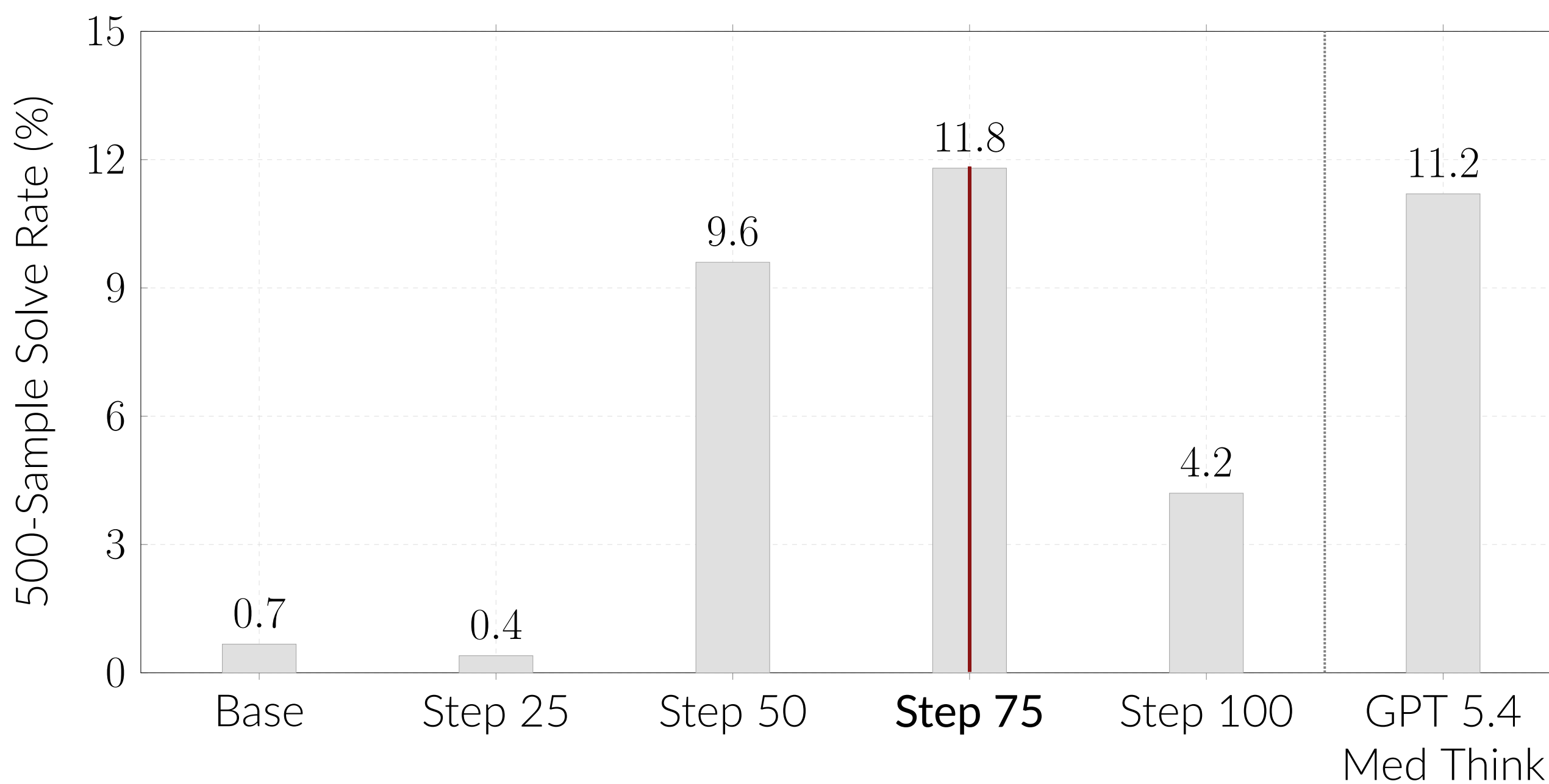
$p < 0.6$: too hard + unreliable labels | $p > 0.8$: too easy + weak gradients
 $0.6 \leq p \leq 0.8$: maximum useful gradient $\times$ label trust

In practice, 10 solutions sampled per subproblem at temperature 0.7. Modal-answer frequency determines retention and provides the training label.

Task: FrontierMath Tier 3 (Arithmetic Geometry)

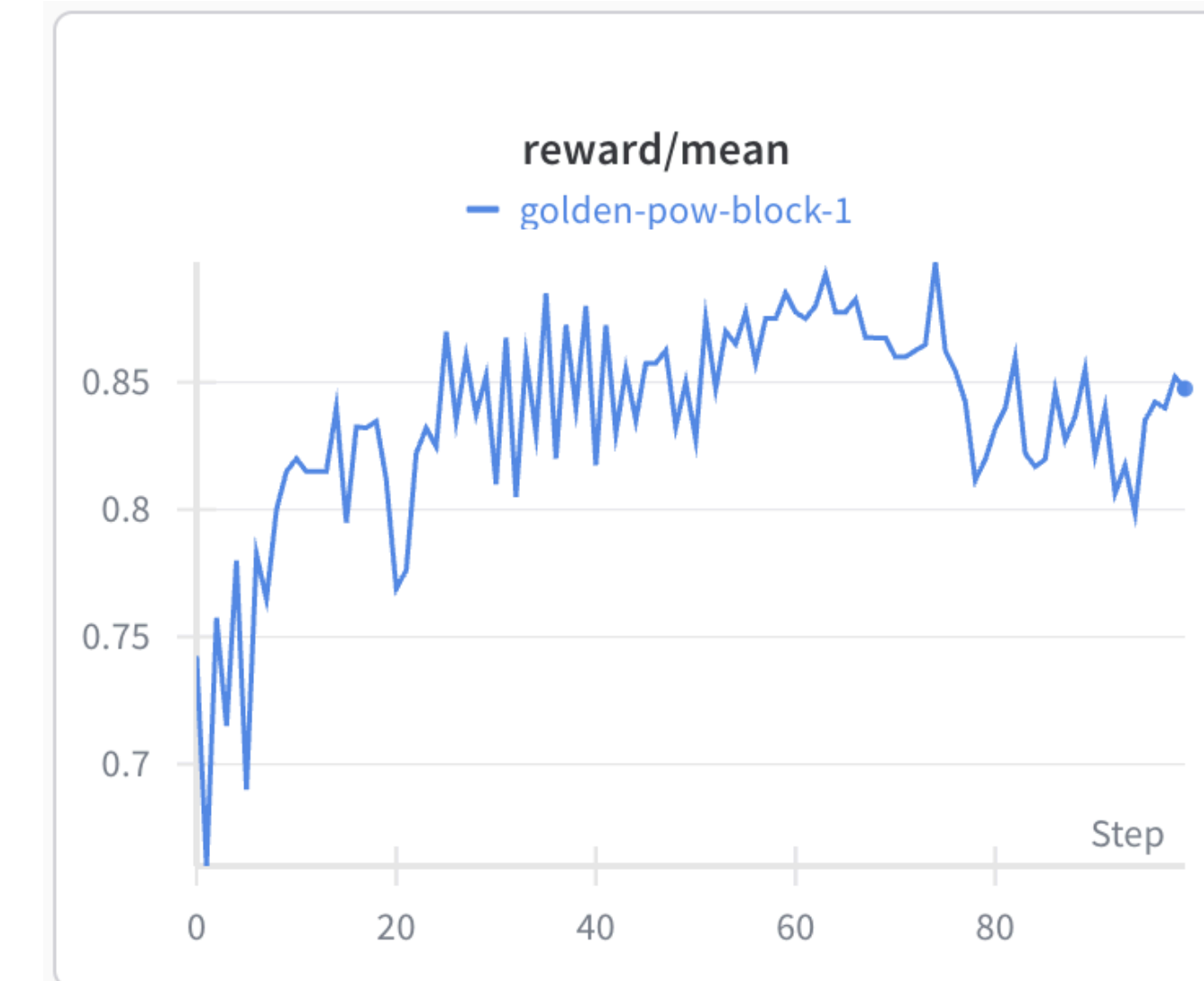
- Tier 3** = early PhD-level; niche tools + novel insights
- Compute $[100L]$ for conics tangent to 5 smooth conics over $\mathbb{F}_p$
- Correct answer: **866** Base model: **4/600 = 0.67%**
- Base model: **gpt-oss-120b**

Results: $\sim 17\times$ Improvement



- Best (step 75): 11.80%** — beats GPT 5.4 Medium Thinking (11.20%)
- Trajectory is **non-monotonic**: drops to 4.20% by step 100

Training Dynamics



- Drop 1 (step ~ 20)**: Solve rate briefly 1.60% then collapses to 0.40%—fragile gains overwritten by noisy labels
- Drop 2 (step ~ 70)**: Solve rate reaches **11.80%** and *persists* at 11.60% through step 80—deep reasoning survives label noise

Pseudo-Label Reliability & Robustness

Validated 50 pseudo-labels against frontier-model majority vote (Claude Opus 4.6, Gemini 3.1 Pro, GPT 5.4 Xhigh). Estimated error rate: **$\sim 50\%$** . Two sources: **(1)** unstable majorities where the correct answer exists but modal vote picks the wrong one, and **(2)** ill-posed subproblems with inconsistent premises that produce disagreement across all models.

Recovery from incorrect supervision:

Trained on pseudo-label **7200** $\rightarrow$ model produces **100** in **10/16** rollouts at step 95
Frontier consensus (GPT 5.4 Med + Claude Opus 4.6) confirms **100** is correct

GRPO learns **intermediate reasoning structure**, not label memorization—the model recovers correct answers despite incorrect supervision.

Conclusions & Future Work

- First curriculum-based TTT for **binary reward** settings
- 0.67% $\rightarrow$ 11.80%** ($\sim 17\times$) on Tier 3 FrontierMath, **no ground-truth labels**
- Main bottleneck: $\sim 50\%$ pseudo-label error rate

Future: formal verification for well-posed subproblems, confidence-based rewards, adaptive checkpoint selection.

References

- [1] Yuksekgonul et al. *Learning to Discover at Test Time*. 2026. [2] Shao et al. *DeepSeekMath*. 2024.
[3] Jiang et al. *VCRL*. 2025. [4] Liu et al. *TTTL*. 2025.