# Coupled Variational Reinforcement Learning for Language Model General Reasoning

***Xueru Wen***, Jie Lou, Yanjiang Liu, Hongyu Lin, Ben He, Xianpei Han, Le Sun, Yaojie Lu, Debing Zhang

*Ph.D. Student, wenxueru2022@iscas.ac.cn*

# Background

## From chain-of-thought to RL w/ Verifiable Reward

1. Things have changed a lot after we realized that we can have LLMs **think "a lot"** before answering.

---

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

---

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2.$

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

$\cdots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

$\sqrt{a - \sqrt{a + x}} = x$

First, let's square both sides:

$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$

Next, I could square both sides again, treating the equation: $\cdots$

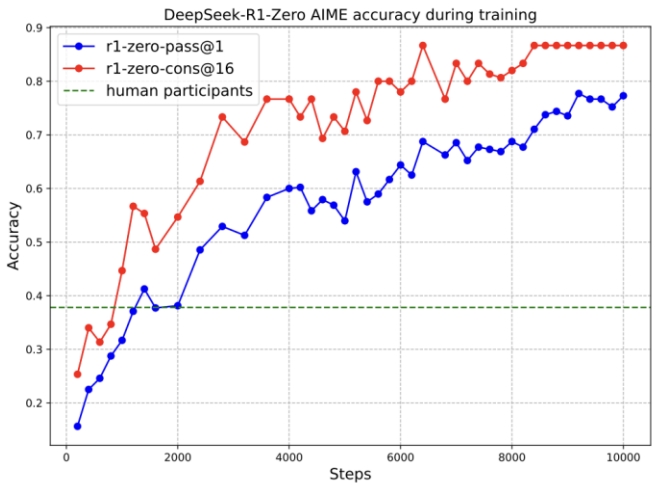
$\cdots$

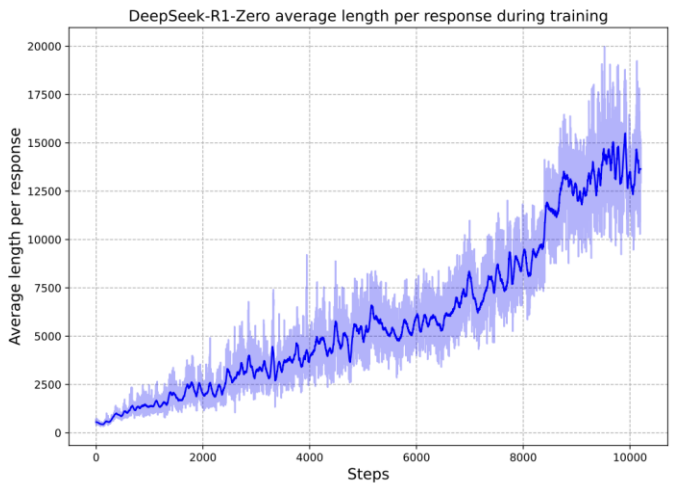
---

# Background

## From chain-of-thought to RL w/ Verifiable Reward

1. Things have changed a lot after we realized that we can have LLMs think "a lot" before answering.
2. This led to a bigger breakthrough: LLMs can learn to think "a lot" **simply through RL with verifiable rewards.**



# Background

## From chain-of-thought to RL w/ Verifiable Reward

1. Things have changed a lot after we realized that we can have LLMs think "a lot" before answering.
2. This led to a bigger breakthrough: LLMs can learn to think "a lot" **simply through RL with verifiable rewards.**

1. Question (x) → LLM Policy →→ Thought + Answer (y)
2. Answer (y) → **Rule-based Verifier (e.g., Test cases/ Math checker)** → Reward (0 or 1)
3. Reward →→ GRPO / PPO update LLM Policy

# Background

## From RLVR to Veri-free RL

1. Though achieving success in math and coding, **highly robust verifiers can be inaccessible** in more general reasoning tasks contrary to what one might expect.

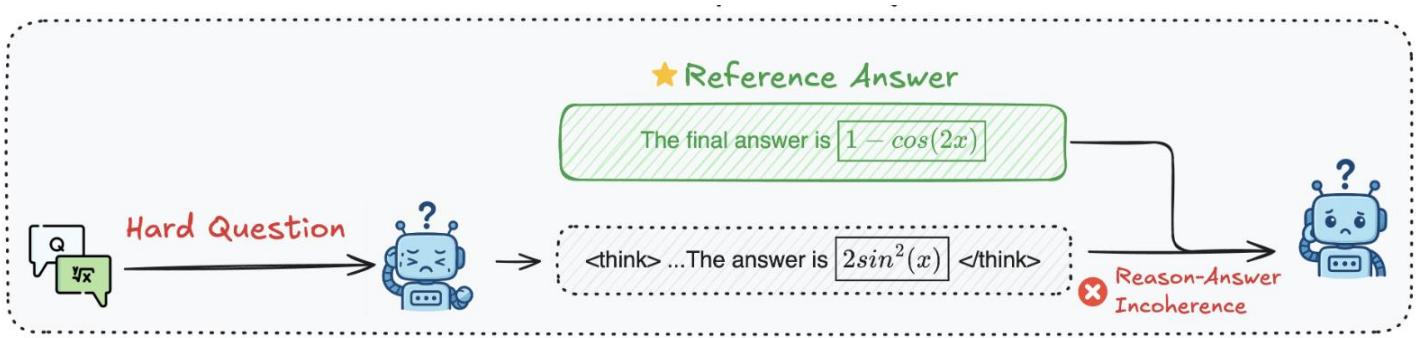
> *"Calcium glycerophosphate contains 1 calcium, 3 carbons, 7 hydrogens, 6 oxygens, and 1 phosphorus. It undergoes double displacement with acetic acid. Knowing calcium is +2 and acetic acid donates only one proton, deduce the final formulas and write the balanced equation."*
>
> The answer is $\underline{2\ CH_3COOH + CaC_3H_7O_6P \rightarrow Ca(CH_3COO)_2 + C_3H_9O_6P.}$

2. The challenge is that **while it may have a short answer, there is no simple rule to verify it**.

# Introduction

## From RLVR to Veri-free RL

1. Though achieving success in math and coding, highly robust verifiers can be inaccessible in more general reasoning tasks contrary to what one might expect.
2. The challenge is that while it may have a short answer, there is no simple rule to verify it.

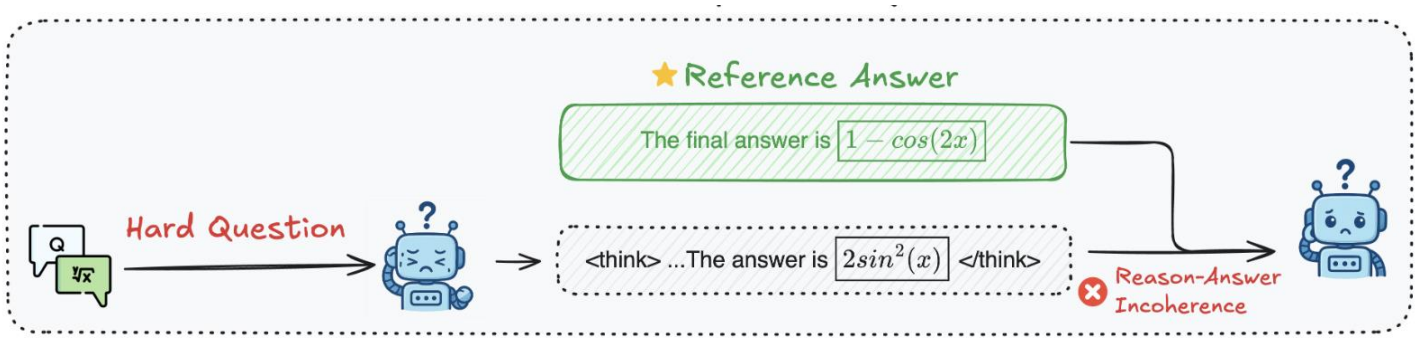



3. Veri-free RL uses the probability of generating the reference answer as the reward, but it suffers from the issue of **reason-answer incoherence**.

# Introduction

## From RLVR to Veri-free RL

1. The challenge is that while it may have a short answer, there is no simple rule to verify it.

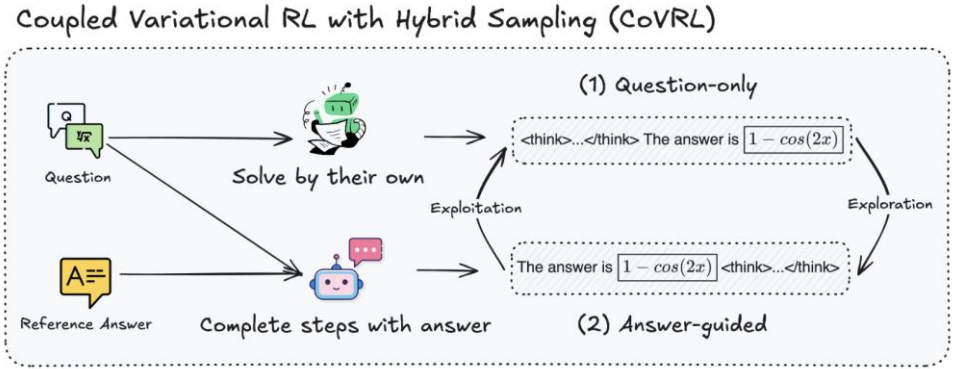



2. Veri-free RL uses the probability of generating the reference answer as the reward, but it suffers from the issue of reason-answer incoherence.
3. Some questions can be too complex for LLMs to solve without guidance, leading to **low sample efficiency**.

# Introduction

## CoVRL (Coupled Variational Reinforcement Learning):

- We unify two generation modes via a hybrid sampling strategy and optimize under a coupled distribution:



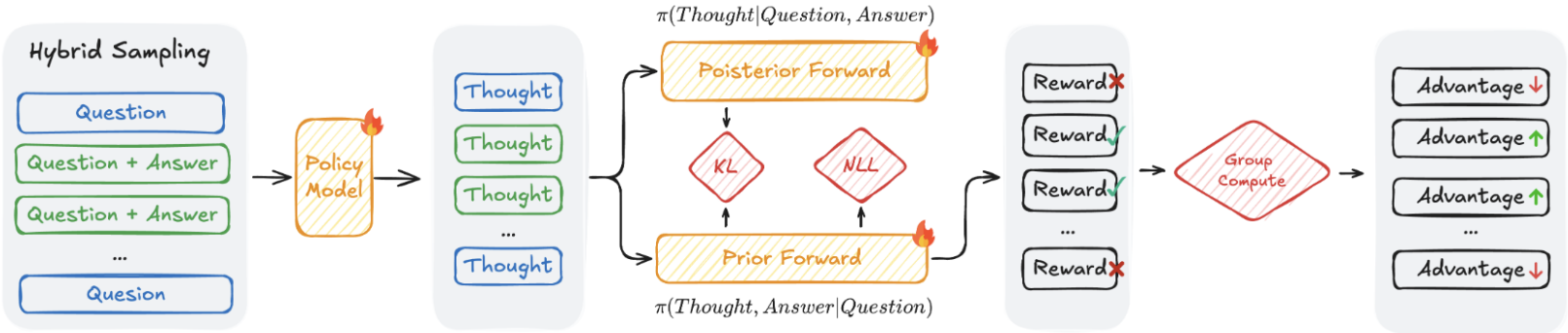

🎯 Answer-Guided Mode (Solves the Pain Points!):

- Fixes Low Efficiency: **Provides the answer as a "hint"**, guiding the model to deduce valid steps without blind searching.
- Fixes Incoherence: Forces the reasoning trace to logically align with the final answer format.

🚀 Question-Only Mode (Preserves Capability):

- Ensures the learned reasoning abilities transfer effectively to real-world inference (where answers are unknown).

# Methodology




**One Model, Two Distributions:**

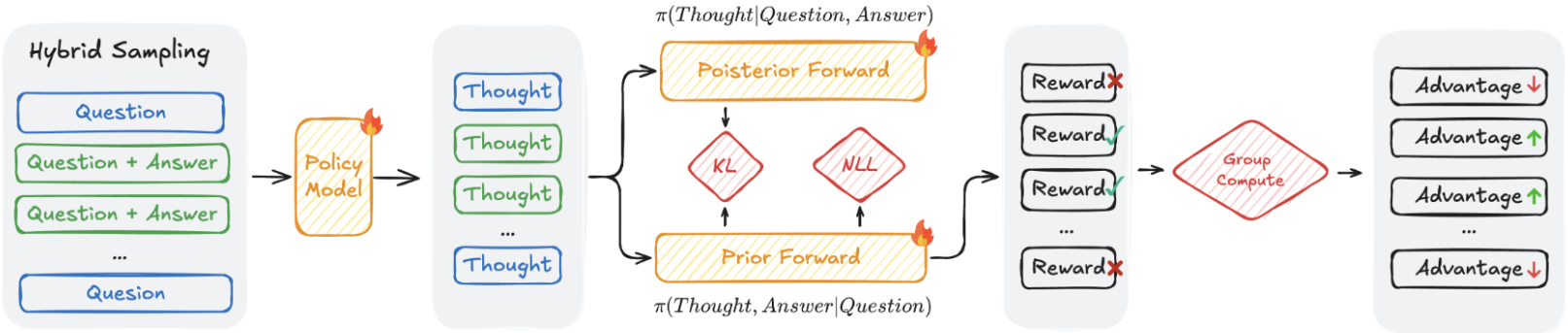
We implement both Prior and Posterior sampling using a single LLM simply by altering the generation order.

- Prior (Question-only): [Q] → <think> → <answer> (Simulates real-world inference).
- Posterior (Answer-guided): [Q + A + <answer>] → <think> (Forces trace-answer coherence).

**Off-Policy Hybrid Sampling:**

- Instead of complex token-level mixing, we randomly sample complete trajectories from either the Prior or Posterior mode.

# Methodology




**The Theoretical Challenge:**

- Randomly mixing sampling trajectories creates a distribution mismatch from the ideal token-level composite distribution.

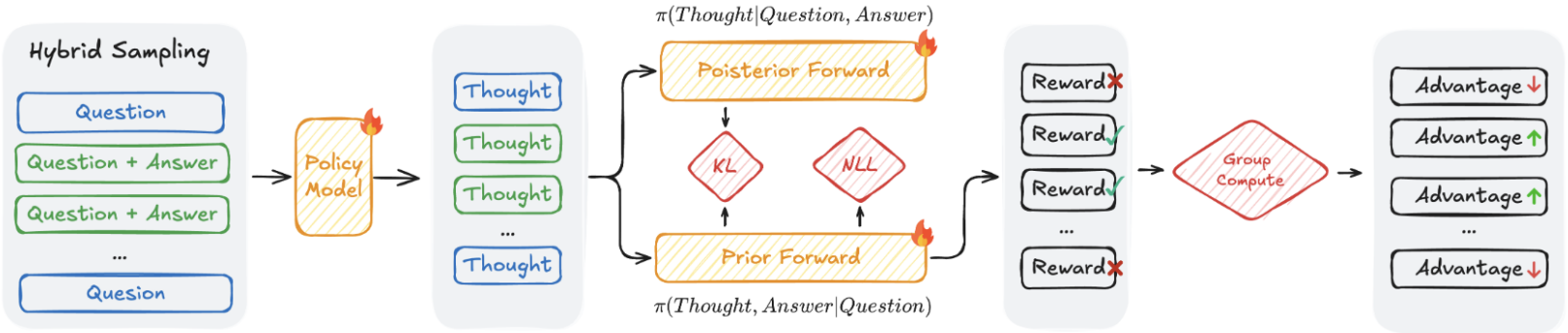
**The Principled Solution (Importance Sampling):**

- We introduce an importance ratio to mathematically correct the gradients, making the optimization theoretically sound.

**Joint Optimization (Optimizing ELBO with Theoretical Guarantees):**

- GRPO (Reward): Evaluates if the reasoning leads to the correct answer. (No external verifier needed!)
- NLL (Positive Reinforcement): Selectively reinforces high-quality traces with positive advantages.
- Bidirectional Dense KL Regularization *(or Self-OPD w/ JSD, may be more well-known):*

# Methodology




**Joint Optimization (Optimizing ELBO with Theoretical Guarantees):**

- GRPO (Reward): Evaluates if the reasoning leads to the correct answer. (No external verifier needed!)
- NLL (Positive Reinforcement): Selectively reinforces high-quality traces with positive advantages.
- Bidirectional Dense KL Regularization *(or Self-OPD, may be more well-known):*
  - 👉 **Prior → Posterior (Anchoring & Exploration):** Constrains the answer-guided trace to not deviate too far from the model's base capability. Guarantees robust transferability to inference.
  - 👈 **Posterior → Prior (Distillation & Exploitation):** Uses the successful, answer-guided paths to pull and update the prior. Acts as step-by-step guidance for efficient exploration.

# Experiment Result

- Our method achieve **a 12.4% improvement** over the base model and **a 2.3% improvement** over the strongest baseline.

| Method | General Reasoning | | | Mathematical Reasoning | | | | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| | GPQA Avg@4 | MMLU-Pro Avg@1 | TheoremQA Avg@2 | AIME'24 Avg@32 | AQuA Avg@4 | CARP-EN Avg@2 | MATH-500 Avg@4 | Minerva Avg@4 | SAT-Math Avg@32 | |
| Base Model | 26.1 | 36.7 | 25.2 | 2.7 | 55.6 | 54.3 | 44.7 | 18.6 | 76.5 | 37.8 |
| VeriFree | 28.9 | 44.1 | 33.4 | 5.0 | 72.6 | 62.8 | 59.5 | 24.0 | 93.3 | 47.1 |
| JLB | **31.6** | 42.7 | 31.9 | 4.8 | 69.5 | 63.4 | 57.6 | 23.6 | 93.7 | 46.5 |
| LaTRO | 31.0 | 42.7 | 32.8 | 4.0 | 67.6 | 50.5 | 59.3 | 24.4 | 90.1 | 44.7 |
| RAVR | 30.2 | 44.5 | 34.8 | 6.3 | 72.4 | 63.2 | 61.2 | 23.3 | 94.1 | 47.8 |
| RLPR | 31.3 | 44.9 | 33.5 | 6.5 | 72.3 | 62.9 | 61.2 | 24.7 | 93.8 | 47.9 |
| **CoVRL (Ours)** | 30.4 | **46.5** | **36.3** | **7.5** | **77.3** | **65.1** | **66.3** | **25.5** | **97.1** | **50.2** |

# Experiment Result

- Our method generalize to **various base model and training dataset**.

*Table 2.* Performance comparison across different base models of various model architectures and sizes.

| Model | General | | | Mathematical | | | | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| | GPQA | MMLU-Pro | TheoremQA | AIME'24 | AQuA | CARP-EN | MATH-500 | Minerva | SAT-Math | |
| | Avg@4 | Avg@1 | Avg@2 | Avg@16 | Avg@4 | Avg@2 | Avg@4 | Avg@4 | Avg@32 | |
| Qwen2.5-7B | 26.1 | 36.7 | 25.2 | 2.7 | 55.6 | 54.3 | 44.7 | 18.6 | 76.5 | 37.8 |
| + CoVRL | 30.4$_{+4.3}$ | 46.5$_{+9.8}$ | 36.3$_{+11.1}$ | 7.5$_{+4.8}$ | 77.3$_{+21.7}$ | 65.1$_{+10.8}$ | 66.3$_{+21.6}$ | 25.5$_{+6.9}$ | 97.1$_{+20.6}$ | 50.2$_{+12.4}$ |
| Qwen2.5-14B | 28.0 | 38.5 | 25.4 | 3.0 | 63.3 | 57.3 | 44.8 | 22.1 | 82.5 | 40.5 |
| + CoVRL | 36.6$_{+8.6}$ | 57.0$_{+18.5}$ | 42.4$_{+17.0}$ | 7.9$_{+4.9}$ | 81.8$_{+18.5}$ | 64.0$_{+6.7}$ | 71.5$_{+26.7}$ | 33.7$_{+11.6}$ | 95.8$_{+13.3}$ | 54.5$_{+14.0}$ |
| Qwen3-8B | 36.9 | 51.0 | 32.8 | 4.0 | 68.6 | 59.4 | 53.6 | 30.0 | 90.5 | 47.4 |
| + CoVRL | 37.6$_{+0.7}$ | 59.6$_{+8.6}$ | 43.7$_{+10.9}$ | 9.2$_{+5.2}$ | 80.5$_{+11.9}$ | 66.1$_{+6.7}$ | 74.5$_{+20.9}$ | 35.2$_{+5.2}$ | 97.6$_{+7.1}$ | 56.0$_{+8.6}$ |
| Qwen3-14B | 37.5 | 57.6 | 37.6 | 7.8 | 75.3 | 64.6 | 67.2 | 33.5 | 93.1 | 52.7 |
| + CoVRL | 42.7$_{+5.2}$ | 63.1$_{+5.5}$ | 46.3$_{+8.7}$ | 9.5$_{+1.7}$ | 83.9$_{+8.6}$ | 65.9$_{+1.3}$ | 76.1$_{+8.9}$ | 38.1$_{+4.6}$ | 97.4$_{+4.3}$ | 58.1$_{+5.4}$ |

*Table 3.* Performance comparison across different training data compositions.

| Training Data | General | | | Mathematical | | | | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| | GPQA | MMLU-Pro | TheoremQA | AIME'24 | AQuA | CARP-EN | MATH-500 | Minerva | SAT-Math | |
| | Avg@4 | Avg@1 | Avg@2 | Avg@16 | Avg@4 | Avg@2 | Avg@4 | Avg@4 | Avg@32 | |
| Base Model | 26.1 | 36.7 | 25.2 | 2.7 | 55.6 | 54.3 | 44.7 | 18.6 | 76.5 | 37.8 |
| Non-Math Only | 30.4$_{+4.3}$ | 46.5$_{+9.8}$ | 36.3$_{+11.1}$ | 7.5$_{+4.8}$ | 77.3$_{+21.7}$ | 65.1$_{+10.8}$ | 66.3$_{+21.6}$ | 25.5$_{+6.9}$ | 97.1$_{+20.6}$ | 50.2$_{+12.4}$ |
| Math Only | 28.9$_{+2.8}$ | 42.7$_{+6.0}$ | 31.8$_{+6.6}$ | 4.2$_{+1.5}$ | 72.1$_{+16.5}$ | 53.1$_{-1.2}$ | 60.1$_{+15.4}$ | 20.6$_{+2.0}$ | 94.3$_{+17.8}$ | 45.3$_{+7.5}$ |

## *Conclusion*

- We introduce **CoVRL** (Coupled Variational Reinforcement Learning).
- You may be interested! If you are doing:

*(1) OPD* *(2) RLVR w/o verifier* *(3) VAE-related idea*

↑ Scan to Read Paper

# THANKS