CORE: Concept-Oriented Reinforcement for Bridging the Definition-Application Gap in Mathematical Reasoning

Zijun Gao[1] Zhikun Xu[2] Xiao Ye[2] Ben Zhou[2]

[1]University of Illinois Urbana–Champaign [2]Arizona State University

Motivation

- **The Illusion of Understanding:** While LLMs excel at complex math, they heavily rely on **procedural pattern matching** rather than genuine conceptual reasoning.
- **The Definition-Application Gap:** Current RL pipelines optimize only for terminal correctness. Consequently, models can **flawlessly memorize theorems but fail to apply them** when superficial surface features (e.g., variables) change.

The Illusion: Perfect Memorization

User: "State the Rational Root Theorem."

LLM: "If p/q is a rational root... then p divides the constant term..."

✅ **Success**

The Reality: Application Failure

The Trap (Question): "If a reduced fraction q/p is a root of f(x), then q divides the constant term."
(Mathematically TRUE, but variables are swapped)

LLM's Reasoning: "It's backwards... p and q are reversed. Reject."

❌ Failed: Superficial Pattern Matching

💡 **Conceptual Fragility:** LLMs master the syntax of definitions but fail the logic of application, collapsing under minor surface perturbations.

Main Research Question: How can we bridge the definition-application gap and explicitly reinforce conceptual reasoning in RL?

Dataset

1. The Seed Corpus

236	703	140
Definitions (C) Curated from the textbook	**Examples** Illustrative content	**Exercises (E)** In-domain dataset

Curated from textbook and translated into English

2. Synthetic Concept Probes (Scale-Up)

Qwen2.5-72B-Instruct Prompted with 236 seed concepts to generate quizzes → **GPT-4o** Quality Control → **1,110 High-Quality Quizzes** For training and diagnosis

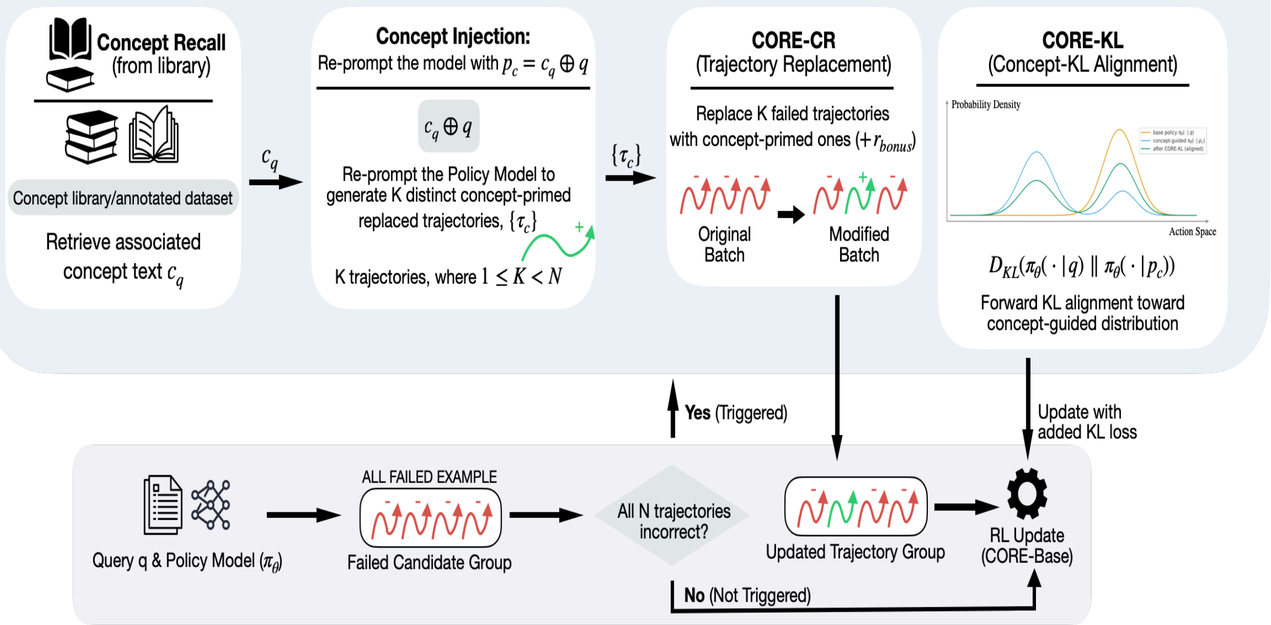
3. Quantitative Gap Diagnostics

Performance crashes under superficial option permutations (Robust Eval)

Qwen2-Math-7B	Standard: 87.0% → Robust: 76%	↓11%
Llama-3-8B	Standard: 70.9% → Robust: 46.8%	↓24.1%

Methodology

Concept-Guided Intervention (CORE) Subsystem



CORE-Base: Standard RL

MECHANISM: Train policy π_θ directly on conceptual quizzes via standard GRPO.

ROLE: Consolidation baseline for implicit concept learning (no runtime guidance)

CORE-CR: Trajectory Replacement

TRIGGER: 0/N correct in a GRPO group

INTERVENTION: Re-prompt with concept c_q to generate K new trajectories, replacing failed ones

REWARD: $R'(\tau_{c,j}) = R(\tau_{c,j}) + r_{bonus}$

CORE-KL: Concept-KL Alignment

TRIGGER: 0/N correct in a GRPO group

INTERVENTION: Distill reasoning from concept-guided "teacher" policy via forward KL-divergence

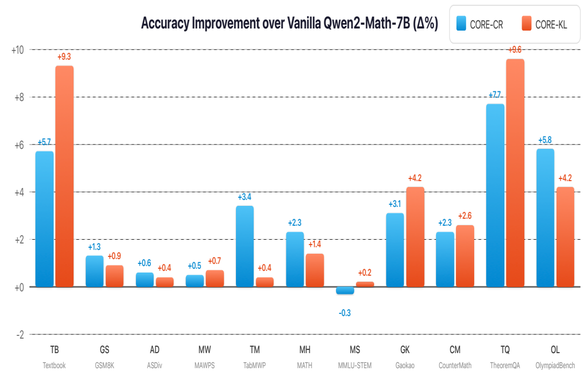
OBJECTIVE:

$$\mathcal{L}_{total} = \mathcal{L}_{GRPO} + \lambda_{KL} \cdot \mathbb{E}[D_{KL}(\pi_\theta(\cdot|p_c)||\pi_\theta(\cdot|q))]$$

Results

Consistent Gains Over Vanilla Baselines

Qwen2-Math-7B: Up to **+9.3%** on in-domain Textbook and **+9.6%** on out-of-domain TheoremQA.



Intrinsic Conceptual Shift & Self-Supervision

Concept-Driven Success Flips

52.6% driven by explicit concept selection

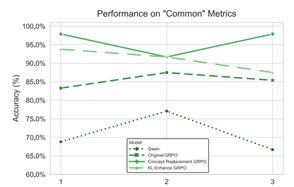
0% relying purely on heuristics

Viability via Self-Supervision (Qwen2-Math-7B-Instruct)

OlympiadBench	GSM8K
+4.2%	+1.6%

Substantial improvements with self-generated quizzes

Robustness to Irrelevant Distractors



Zero Architectural Changes

Plug-and-play improvements across both base and instruction-tuned models

DeepSeek-R1-DQ-1.5B	MMLU-STEM	+1.3%
Qwen2.5-Math-1.5B	Minerva Math	+3.3%
Llama-3-8B-Instruct	TabMWP	+3.3%