CORE: CONCEPT-ORIENTED REINFORCEMENT FOR BRIDGING THE DEFINITION–APPLICATION GAP IN MATHEMATICAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) often solve challenging math exercises yet fail to apply the concept right when the problem requires genuine understanding. Popular Reinforcement Learning with Verifiable Rewards (RLVR) pipelines reinforce final answers but provide little fine-grained conceptual signal, so models improve at pattern reuse rather than conceptual applications. We introduce CORE (Concept-Oriented REinforcement), an RL training framework that turns explicit concepts into a controllable supervision signal. Starting from a high-quality, low-contamination textbook resource that links verifiable exercises to concise concept descriptions, we run a sanity probe showing LLMs can restate definitions but fail concept-linked quizzes, quantifying the conceptual reasoning gap. CORE then (i) synthesizes additional concept-aligned quizzes, (ii) injects concept snippets into rollouts, and (iii) reinforces the conceptual reasoning by replacing with correctly concept-applied trajectories or constraining drift with a lightweight divergence penalty; the procedure is compatible with standard policy-gradient methods. On two 7B models, CORE yields consistent gains over the vanilla baseline and SFT training across in-domain concept–exercise suites and diverse out-of-domain math benchmarks. CORE demonstrates that concept-injected, outcome-regularized rollouts supply the missing fine-grained supervision needed to bridge question-solving competence and true conceptual reasoning without committing to a particular RL algorithm or certain process-based verifiers.

1 INTRODUCTION

Recent LLMs are becoming good at tackling competition-level questions, yet they fall short of conceptual math reasoning beyond applying competition tricks or executing numerical calculations (Yang et al., 2024b; Guo et al., 2025a; Huang & Yang, 2025; Chen et al., 2025). Here, conceptual reasoning means identifying the right concept and applying it in the solution, as opposed to procedural pattern matching often sufficient for GSM8K (Cobbe et al., 2021) or MATH (Hendrycks et al., 2021b) and exposed by perturbation-based tests (Patel et al., 2021; Yu et al., 2024; Mirzadeh et al., 2025; Huang et al., 2025) On many benchmarks, models can mimic solution templates, chain together routine algebraic steps, and even memorize recurring patterns—while still choosing the wrong concept for a problem or failing to correctly apply certain concepts. This gap matters: solving a word problem by spotting a familiar cue is not the same as understanding linear independence, continuity, or convexity and deploying those notions correctly (Li et al., 2025; Huang et al., 2025).

Two main factors are contributing to this gap. First, large collections of exercise-style problems can often be solved by exploiting surface regularities (formats, keywards, and step patterns) rather than engaging the intended mathematical concepts (Guo et al., 2025b; Wu et al., 2025). Second, widely used RLVR pipelines (Schulman et al., 2017; Shao et al., 2024) optimize a terminal scalar reward for correctness. Such signals can improve search heuristics (Qin et al., 2025) but are too coarse to specify which concept should be invoked, where it should enter the argument, or how it supports subsequent steps. Moreover, directly assessing conceptual understanding is also challenging: concepts cannot be instantiated without a clear motivation or goal. We operationalize this by pairing concise concept definitions with aligned quiz questions that necessitate the concept. In a preliminary diagnostic, models readily recites the concepts yet frequently fail the quizzes, revealing a pronounced definition-application gap.

In order to mitigate this definition-application gap, we propose `CORE` (Concept-Oriented REinforcement), an RL-based framework that turns explicit mathematical concepts into concept-driven training signals in sampling. `CORE` starts from curating a high-quality data that (i) provide human-verified exercises and (ii) link each exercise to the underlying concept(s), which serve as the in-domain test set and seed data for further generation and training signals. We then expand coverage by generating additional concept-aligned quizzes using strong LLMs to curate the training set. For the training recipe, `CORE` has explored three designs: ❶ **Original RL (`CORE`-Base):** directly training with the generated quizzes by RL algorithms, ❷ **Concept Enhancement (`CORE`-CR):** injecting concise concept snippets into rollouts to replacing half of original ones if all trajectories are incorrect, and ❸ **KL Divergence (`CORE`-KL):** implementing the KL divergence term between the concept-guided trajectories and the original ones to implicitly constraint the model towards using concepts. The above design choices investigate in three main components, **data**, **rollouts**, and **loss function**, of RL algorithms to mitigate the definition-application gap and improve the conceptual reasoning. Moreover, this framework wraps around standard policy-gradient reinforcement algorithms without architectural changes. At test time, the trained model is evaluated without providing the concept text, measuring whether concept-aware training translates into genuine conceptual competence rather than reasoning shortcuts.

Empirical results show that `CORE` delivers consistent and significant improvements across both **Qwen2-Math-7B** and **Llama-3-8B-Instruct**. On Qwen2-Math-7B, CORE variants yield improvements of up to +14.8%, while on Llama-3-8B-Instruct, CORE surpasses Vanilla by up to +3.2% across benchmarks. These results demonstrate that, without any architectural modifications, `CORE` can effectively enhance conceptual application and reasoning ability through explicit concept injection combined with concept-aware sampling and divergence control.

2 Related Works

Mathematical Reasoning in LLMs Recent systems approach math reasoning through specialized training or sheer scale. WizardMath (Luo et al., 2025) uses reinforcement learning from Evol-Instruct feedback. MAmmoTH (Yue et al., 2024) blends chain-of-thought and program-of-thought for hybrid tuning. Qwen2.5-Math (Yang et al., 2024b) continues from a general model on a large curated math corpus. Llemma (Azerbayev et al., 2024) is pre-trained on Proof-Pile. DeepSeekMath (Shao et al., 2024) adds about 120B math tokens and introduces *Group Relative Policy Optimization* (GRPO). InternLM2-Math (Ying et al., 2024) unifies chain-of-thought, reward modeling, and formal reasoning. General-purpose models also lean heavily into math: Llama 3.1 (Team, 2024) up-samples math data, DeepSeek-R1 and DeepSeek-V3 (DeepSeek-AI, 2025a;b) raise math and programming proportions across trillions of tokens, Claude 3.7 (Anthropic, 2025) emphasizes transparent multi-step reasoning, Gemini (Gemini, 2025) targets long-chain deduction, and OpenAI o1 and o3 (OpenAI, 2024; 2025) scale test-time compute. Despite these advances, many pipelines reward final answers or rely on data scale, leaving concept selection and application under-taught. Our `CORE` addresses this by injecting explicit concept signals into rollouts and regularizing outcomes, yielding consistent gains on concept-dependent evaluations while remaining agnostic to the underlying RL algorithm.

Mathematical Benchmarks A wide range of math benchmarks now probes reasoning at various levels. At the elementary level, GSM8k is the standard for multi-step math problems. MAWPS (Koncel-Kedziorski et al., 2016) aggregates sources, ASDiv (Miao et al., 2021) adds type and grade labels, and SVAMP (Patel et al., 2021) stresses robustness through controlled perturbations. Cross-lingual coverage includes CMATH for Chinese primary school (Wei et al., 2023) and CN Middle School 24, while standardized suites such as Gaokao 2023 EN, SAT Math, and MMLU-STEM enable broader STEM-wide comparisons (Zhong et al., 2023; Hendrycks et al., 2021b). For competition-level reasoning, MATH curates Olympiad and contest problems with stepwise solutions, and OlympiadBench (He et al., 2024) extends to bilingual and multimodal settings that emphasize proof-style reasoning and reduce contamination risk. Our evaluations of `CORE` are based on in-domain concept–exercise suites and out-of-domain math benchmarks, showing consistent gains over strong baselines and highlighting improvements specifically on concept-dependent categories.

Conceptual Reasoning Answer accuracy alone doesn't reveal whether the right concepts were selected and used, several works probe whether models actually select and use the right concepts. Specifically in math, conceptual reasoning requires people to reason around math concepts and axioms at the play of math hypothesis, statements and problems (Simon, 2011). THEOREMQA (Chen et al., 2023) targets theorem application across STEM, explicitly requiring mapping from a named theorem to its correct use in problem solving. GSM-SYMBOLIC (Mirzadeh et al., 2025) examines symbolic generalization limits of models trained on GSM8K-style data. COUNTERMATH (Li et al., 2025) proposes counterexample-driven, concept-sensitive evaluations to diagnose superficial cues versus true concept use. Complementary directions include a conceptualization framework that maps abstract questions into verifiable symbolic programs (Zhou et al., 2024), a self-supervised analogical learning scheme that transfers high-level solutions across cases (Zhou et al., 2025), and a Bayesian inference formulation coupling abductive proposals with structured deduction for calibrated decisions (Feng et al., 2025). Overall, these efforts indicate that a single end-point score is insufficient for assessing whether models select and correctly apply concepts; concept-aligned training signals or structured evaluation protocols are required.

3 CORE: CONCEPT-ORIENTED REINFORCEMENT LEARNING

3.1 OVERVIEW

We study math reasoning where success depends mainly on conceptual math reasoning rather than replaying surface templates in the training data. Our proposed framework CORE has been developed through the following stages: **dataset curation**, **gap diagnostics**, and **concept reinforcement recipe**. For *dataset curation*, we have leveraged a classical mathematical textbook with clear associations between concepts and exercises for training and evaluation. For *gap diagnostics*, we have used the curated data to both qualitatively and quantitatively identify the definition-application gaps in the conceptual mathematical reasoning. For *concept reinforcement recipe*, we have mainly designed three training recipes in reinforcing the models' conceptual mathematical reasoning.

3.2 DATASET CURATION

To acquire rigorous conceptual reasoning signals, we curated a corpus from a canonical textbook, *Advanced Algebra* (3rd Edition) (Yao & Xie, 2015). This source was chosen for two-fold reasons. First, it provides a comprehensive and structured curriculum in linear algebra, progressing logically from foundational concepts like determinants and matrices to advanced topics such as linear spaces and canonical forms. Its ten chapters are methodically structured, each containing: i) core concept definitions ($\mathcal{C}$), ii) illustrative examples, and iii) concept-aligned exercises ($\mathcal{E}$), where exercises in a given chapter primarily test the concepts introduced in that same chapter. The textbook's long-standing use and human verification ensure a logical progression of topics and coherent conceptual dependencies, making it an ideal corpus for developing and evaluating a concept-oriented learning paradigm. Second, by manually translating this Chinese textbook into English, we significantly mitigate the risk of training data contamination present in many existing English-language corpora.

The extraction yielded 236 concept texts, 703 examples, and 140 multiple-choice questions sourced from the exercise sections. More details are provided in Appendix A.

3.3 GAP DIAGNOSTICS

3.3.1 PROBING THE GAP BETWEEN KNOWLEDGE RECITATION AND APPLICATION

The structured nature of our curated corpus, with its explicit mapping between concepts ($\mathcal{C}$) and exercises ($\mathcal{E}$), provides an ideal testbed to diagnose a critical failure mode in state-of-the-art LLMs: the disconnect between parametric knowledge and its application in problem-solving. We conduct the following sanity check to probe this gap, which provides the foundational motivation for our proposed training framework.

We perform a qualitative analysis on the multiple-choice exercises that a powerful baseline, GPT-4o[1], failed due to conceptual errors. After a model fails an exercise, we prompt it to describe the core

[1] `https://openai.com/index/hello-gpt-4o/`

User
Question: Let $f(x)$ be a polynomial with integer coefficients. Which of the following statements is true? A) $f(x)$ has a rational root iff it is reducible over the rationals. B) If a reduced fraction $\frac{q}{p}$ is a root of $f(x)$, then q divides the constant term of $f(x)$. C) If a prime p divides all coefficients of $f(x)$ except the leading term, then $f(x)$ is irreducible. D) If $f(x)$ has a repeated factor, then it must have a repeated root. Solve it step by step.

Model
Statement B: "If a reduced fraction $\frac{q}{p}$ is a root of $f(x)$, then q divides the constant term."
False. This misstates the *Rational Root Theorem*. If $r = \frac{q}{p}$ is a rational root in lowest terms, then $q \mid$ constant term and $p \mid$ leading coefficient.
So it's backwards: the **numerator** divides the constant term, not the denominator. This statement is false.

User
Please state Rational Root Theorem.

Model
Rational Root Theorem: For $f(x) = a_n x^n + \cdots + a_0$ with integer coefficients $a_i \in \mathbb{Z}$. If $\frac{p}{q}$ is a rational root of $f(x)$ in lowest terms, then $p \mid a_0$ and $q \mid a_n$.

User
Use Rational Root Theorem to solve the question again.

Model
Option B reverses the roles of p and q, so it is **wrong**. Reject B.

Figure 1: An example of ChatGPT-4o's superficial understanding of the Rational Root Theorem. Please read from left to right.

mathematical concept associated with that problem (e.g., "Describe the concept of Linear Independence."). We observe that in the majority of cases, the model can accurately and comprehensively recite the correct definition and properties of the concept, indicating that the requisite knowledge is parametrically encoded. A classic example is shown in Figure 1. GPT-4o is able to correctly restate the Rational Root Theorem. However, when the numerator and denominator in the problem are swapped, the model still follows its ingrained reasoning and makes an incorrect judgment.

These sanity checks reveals a critical failure mode: the model's problem-solving process appears to be locked into a rigid, pattern-matching heuristic, failing to flexibly incorporate or be guided by explicit conceptual knowledge, even when it is readily available. This observed gap between the ability to recite a concept and the ability to apply it motivates the need for a training paradigm that explicitly forces the model to ground its reasoning in concepts, which we introduce in the following section.

3.3.2 Synthetic Concept Probes and Robustness Evaluation

Measuring Conceptual Understanding via Concept Probes. A fundamental challenge in evaluating mathematical reasoning is the difficulty of quantitatively measuring a model's grasp of specific concepts. High-level benchmark scores often obscure fine-grained conceptual failures. The highly structured nature of our curated textbook, however, provides a unique opportunity to address this challenge. To create a direct and quantifiable measure of conceptual understanding, we introduce the idea of **Concept Probes**: targeted quizzes generated directly from the textbook's concept definitions and theorems, where a model's performance on these probes serves as a proxy for its mastery of the underlying concepts.

To realize this, we constructed a new dataset of conceptual quizzes. This process involved two key stages: generation and validation.

1. **Generation:** We prompted a powerful generator model, Qwen2.5-72B-Instruct, to create 5–8 multiple-choice quizzes for each of the 236 concept texts in our corpus. This resulted in a candidate pool of 1,200 quizzes, each designed to be closely tied to its source concept and formatted with standard LaTeX.

2. **Validation:** To ensure the quality and validity of these synthetic quizzes, we designed a rigorous filtering pipeline using a separate, powerful assessor model, GPT-4o. This cross-model validation strategy is intentionally designed to reduce harvester bias. For each quiz, GPT-4o evaluated six dimensions (e.g., clarity, correctness, uniqueness) and provided an overall rating and a confidence level. We discarded the 90 quizzes that were rated "Fair" or "Poor" with high confidence. This stringent process yielded a final set of **1,110 high-quality quizzes** that serve as our Concept Probes.

Diagnostic Experiment: Robustness to Superficial Perturbations. To validate our hypothesis that models rely on superficial heuristics, we designed a diagnostic experiment to test the robustness of their conceptual knowledge. Using our 1,110 curated quizzes, we use a **Robust Evaluation** protocol. For each quiz, we generate three variants by randomly permuting the order of its multiple-choice options. A model is considered to have *robustly* solved a problem **only if** it correctly answers the original question *and* all three of its permuted variants. This protocol is designed to test whether a model's understanding is invariant to semantically-irrelevant changes that preserve the core concept.

We applied both the standard and our Robust Evaluation protocols to a suite of contemporary models, including Qwen2-Math-7B, OLMo-7B-Instruct, and Llama-3-8B. The results, illustrated in Table 1, reveal a stark and consistent performance gap across all models. For instance, while a model like OLMo-2-7B may achieve high accuracy (e.g., $> 70\%$) under the standard protocol, its performance plummets to below 50% under Robust Evaluation. This significant degradation provides strong empirical evidence for our hypothesis, demonstrating that the models' success is heavily reliant on shallow heuristics rather than a deep, structural understanding of the underlying concepts.

Table 1: Performance comparison under Original vs. Robust Evaluation protocol across different models. The best performance in each column is highlighted in bold.

	Standard evaluation		Robust evaluation	
Model	**pass@1 accuracy**	**self-consistent**	**pass@1 accuracy**	**self-consistent**
Qwen-2-Math-7B	74.33% +0	87.0% +0	45.92% -28.4	76.0% -11.0
OLMo-2-7B	57.83% +0	70.17% +0	36.25% -21.6	44.42% -25.8
LLaMA-3-8B	44.75% +0	70.92% +0	20.25% -24.5	46.75% -24.2

3.4 Concept Reinforcement Recipe

To bridge the gap between procedural mimicry and conceptual understanding, we propose CORE. CORE is an RL-based framework designed to inject conceptual knowledge into models through any policy gradient based RL algorithm. The core idea is to conditionally intervene during training with concept-guided instruction precisely when the model demonstrates a failure in understanding, guiding the policy update towards a more robust, concept-grounded reasoning process. Our CORE framework mainly consists of the following three design choices based on the standard GRPO.

CORE-Base: Standard RL on Conceptual Quizzes. The foundational approach within our framework, **CORE-Base**, involves training the policy π_θ directly on our curated set of conceptual quizzes ($\mathcal{Q}$) using the standard GRPO algorithm. In this setting, the model learns from the conceptual data without any further explicit guidance during the training process. This approach measures the model's ability to implicitly learn concepts from the rich question-answer pairs generated from concepts.

CORE-CR: *C*oncept-Guided Trajectory *R*eplacement. Building upon the base setting, **CORE-CR** introduces a conditional intervention triggered by a *conceptual failure event* (i.e., all N responses in a GRPO group are incorrect). Upon triggering, we form a concept-guided prompt $p_c = c_q \oplus q$ where q is the original problem from our quiz dataset, c_q is its associated ground-truth concept text, and $\oplus$ denotes concatenation. We then generate K new trajectories $\{\tau_{c,1}, \ldots, \tau_{c,K}\}$ from the concept-guided policy, where $1 \leq K < N$. We then **randomly select and replace** K trajectories from the original failed group with these new concept-guided ones.

To incentivize learning with concepts, we assign the new trajectories an augmented reward:

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

where $r_{\text{bonus}} > 0$ is a hyperparameter. The GRPO update is then performed on this partially replaced, concept-guided batch. Notably, there is a recent work called BREAD (Zhang et al., 2025), which shares a very similar methodology to CORE-CR while derives from rethinking the advantages of SFT and RL instead of improving conceptual reasoning.

CORE-KL: Concept-Guided *KL*-Regularization. The **CORE-KL** method introduces a fine-grained regularization signal to guide the policy's internal reasoning process. This approach is also triggered by a *conceptual failure event*. Instead of directly replacing trajectories, this method encourages the model's standard step-by-step predictive process at each timestep t, denoted $\pi_\theta(\cdot \mid q, y_{<t})$, to align with the more robust process it exhibits when primed with a concept, $\pi_\theta(\cdot \mid p_c, y_{<t})$.

We formulate this as a **forward KL-divergence** objective. This choice is deliberate: it encourages the base policy to cover the full distribution of reasoning paths considered by the concept-guided "teacher" policy, rather than collapsing to a single mode, fostering a more comprehensive distillation of the entire reasoning process. Upon a conceptual failure trigger, we first sample a high-quality reference trajectory, $Y^* = (y_1^*, \ldots, y_T^*)$, from the *current, online concept-guided policy*, i.e., $Y^* \sim \pi_\theta(\cdot \mid p_c)$. Our objective is then to minimize the KL-divergence between the next-token predictive distributions of the guided and un-guided policies at each timestep t, conditioned on the prefix of the reference trajectory $Y^*_{<t}$. This is formulated as a loss term added to the base RL objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GRPO}} + \lambda_{\text{KL}} \cdot \mathbb{E}_{Y^* \sim \pi_\theta(\cdot|p_c)} \left[\sum_{t=1}^{|Y^*|} D_{\text{KL}}\Big(\pi_\theta(\cdot \mid p_c, y^*_{<t}) \,\big\|\, \pi_\theta(\cdot \mid q, y^*_{<t})\Big) \right] \tag{1}$$

where $\pi_\theta(\cdot \mid \text{context}, y^*_{<t})$ is the current policy's probability distribution over the next token. This forces the model's internal reasoning process on the original problem q to faithfully mimic the process it would follow if it were explicitly given the concept c_q.

4 Experiments

4.1 Baseline Models

We select **Qwen2-Math-7B** (Yang et al., 2024a) as our primary evaluation model. This choice is motivated by its moderate mathematical proficiency: it demonstrates non-trivial reasoning skills while still leaving sufficient headroom on our quiz tasks, thus offering a reliable foundation for assessing the impact of CORE . Beyond this, we further show that the CORE algorithm generalizes effectively to **Llama-3-8B-Instruct** (Team, 2024), highlighting its robustness and applicability to instruction-tuned models as well.

4.2 Training Settings

Our CORE interventions introduce method-specific hyperparameters that are applied on top of the CORE-base, i.e., original GRPO with generated quizzes, setup. For the **Concept-Guided Trajectory Replacement** method, concept-guided trajectories receive a reward bonus of $r_{\text{bonus}} = 0.4$. For the **Concept-Guided KL-Regularization** method, an **additional and dynamic** KL coefficient, λ_{KL}, is applied: we use $\lambda_{\text{KL}} = 0.03$ if the reference concept-guided trajectory is correct, and $\lambda_{\text{KL}} = 0.005$ if it is incorrect. A comprehensive list of all other hyperparameters, derived from our training script, is provided in Appendix B.1..

4.3 Evaluation Settings

In-domain Test. To measure in-domain performance, we use the 140 multiple-choice exercises curated from the textbook. These exercises serve as a high-quality and reliable measure of concept application due to their expert authorship and direct alignment with the textbook's definitions. We denote this test set as **Textbook**.

Out-of-domain Benchmarks. To assess whether the conceptual understanding fostered by CORE generalizes beyond our curated training data, we evaluate our trained models on a diverse suite of out-of-domain benchmarks. These benchmarks were specifically chosen to probe for different facets of mathematical reasoning, from multi-step arithmetic and competition math to robustness against perturbations. This evaluation is critical to demonstrate that CORE does not simply overfit to the textbook's style, but rather instills a more fundamental and transferable reasoning capability.

We evaluate models trained with three instantiations of our CORE framework—CORE-Base, CORE-CR, and CORE-KL—on the following out-of-distribution benchmarks: **GSM8K** (Cobbe et al., 2021), **ASDiv** (Miao et al., 2021), **MAWPS** (Koncel-Kedziorski et al., 2016), **TabMWP** (Lu et al., 2023), **MATH** (Hendrycks et al., 2021b), **MMLU-STEM** (Hendrycks et al., 2021a), **Gaokao 2023 (EN)** (Zhong et al., 2023), **CounterMath** (Li et al., 2025), **TheoremQA** (Chen et al., 2023), and **OlympiadBench** (He et al., 2024). A detailed description of the datasets is provided in Appendix B.2.

Metrics. Across all benchmark evaluations in this paper, we employ a self-consistency protocol to ensure robust and stable results. For each problem, we sample 21 distinct reasoning paths by setting a high sampling temperature ($T = 0.7$). We denote this as **SC@21**. More evaluation details are shown in the Appendix B.3

Table 2: Main table of accuracy (%) under SC@21 (T=0.7). Columns use *two-letter* abbreviations: **TB**=Textbook, **GS**=GSM8K, **AD**=ASDiv, **MW**=MAWPS, **TM**=TabMWP, **MH**=MATH, **MS**=MMLU-STEM, **GK**=Gaokao 2023 (EN), **CM**=CounterMath (reported as F1), **TQ**=TheoremQA, **OL**=OlympiadBench.

Model	Method	TB	GS	AD	MW	TM	MH	MS	GK	CM	TQ	OL
Qwen2-Math-7B	Vanilla	46.4	89.8	95.1	96.8	90.2	69.1	72.9	55.3	13.2	34.6	28.7
	SFT	45.0	86.6	94.1	96.6	85.6	59.4	72.4	46.5	16.7	**44.2**	19.7
	CORE-Base	50.7	90.8	95.4	97.2	92.6	**71.7**	72.9	**59.5**	13.5	40.4	33.9
	CORE-CR	52.1	**91.1**	**95.7**	97.3	**93.6**	71.4	72.6	58.4	15.5	42.3	**34.5**
	CORE-KL	**55.7**	90.7	95.5	**97.5**	90.6	70.5	**73.1**	**59.5**	**15.8**	**44.2**	32.9

As shown in Table 2, models trained with the CORE framework exhibit consistent and significant performance improvements across the majority of out-of-domain benchmarks when compared to the vanilla baseline. This demonstrates that CORE successfully enhances the models' underlying reasoning abilities in a way that generalizes to unseen problem distributions and formats.

Table 3: Llama-3-8B-Instruct: Out-of-domain benchmark accuracy (%) under SC@21 (T=0.7). Columns use *two-letter* abbreviations: **GS**=GSM8K, **AD**=ASDiv, **MW**=MAWPS, **TM**=TabMWP, **MH**=MATH, **MS**=MMLU-STEM, **GK**=Gaokao 2023 (EN), **OL**=Olympiad. For each dataset, we use the first $\min(0.2N, 500)$ questions as the evaluation subset, where N is the total number of questions in the dataset.

Model	Method	GS	AD	MW	TM	MH	MS	GK	OL
Llama-3-8B-Inst	Vanilla	93.9	91.0	82.5	**65.5**	75.0	**26.0**	65.6	**11.1**
	CORE-CR	**95.9**	**93.7**	**84.5**	61.8	75.0	24.7	**68.8**	7.4
Qwen2-Math-7B	Vanilla	91.6	96.6	98.3	95.0	83.4	**68.6**	**52.0**	12.6
	CORE-CR	**93.1**	**97.1**	**98.6**	**95.5**	**85.2**	67.6	50.7	**14.8**

5 Analysis

5.1 Does CORE Enhance Concept Selection and Application?

We first verify what training with CORE actually changed: are the observed accuracy gains attributable to improved *concept selection and application*, rather than superficial heuristics, and achieved without any test-time concept prompting? To probe this, we evaluate four models on 140 textbook exercises—Vanilla, CORE-Base, CORE-CR, and CORE-KL—and construct a diagnostic subset W consisting of problems on which either Vanilla or CORE-BASE fails while both CORE variants succeed,

$$W = \{\, i \mid (\text{Vanilla fails on } i \text{ or CORE-BASE fails on } i) \wedge (\text{CORE-CR and CORE-KL succeed on } i) \,\},$$

Category	# Problems	Probability(%)
Concept-selection	10	52.6
Mixed	9	47.4
Heuristic-selection	0	0.0
Total	19	100.0

Table 4: Results on the diagnostic subset W ($|W| = 19$). A problem is Concept-Selection iff both CORE-CR and CORE-KL explicitly invoke the target concept and show no heuristic cues; Heuristic-selection iff both rely on heuristics with no substantive concept use; otherwise Mixed.

yielding $|W| = 19$ (12 Vanilla-only failures, 4 CORE-BASE-only failures, 3 shared failures). For each $i \in W$, we read the generations from CORE-CR and CORE-KL and score along two dimensions: (i) **concept hits**—the output explicitly mentions the task's target concept and uses it correctly in the reasoning; and (ii) **heuristic cues**—guessing, option elimination without justification, surface pattern matching, or plug-in substitution without conceptual warrant.

For labeling, we keep the rules simple. A problem is called Concept-Selection if both CORE outputs (CORE-CR and CORE-KL) contain a concept hit and neither shows heuristic cues; it is Heuristic-Selection if both rely on heuristics and show at most one concept hit; otherwise it is Concept+Heuristic (mixed). We require "two hits" (one per CORE variant) so that if either variant omits the concept, the instance is not counted as Concept-Selection. As Table 4 shows, 10/19 (52.6%) cases are Concept-Selection , 9/19 (47.4%) are Mixed, and 0/19 are Heuristic-selection. Due to our strict constraints, it rules out superficial shortcutting as the primary driver of the gains. On one representative question (see Table 5), even after appending a targeted concept prompt at test time, CORE-Base remained incorrect, whereas CORE-KL solved it without any prompt. While this is a single illustrative case, it reinforces that the observed gains come from training-induced mechanism change rather than prompt engineering. Taken together, the evidence indicates a mechanism shift: CORE improves accuracy mainly by strengthening concept selection and application.

5.2 Does CORE Improve Robustness to Irrelevant Concept Perturbations?

We next ask whether training with CORE yields improved *robustness* to irrelevant concept cues. Using 140 high-quality, thematically related textbook exercises, we prepend $K \in \{1, 2, 3\}$ concepts that are not directly related to the target concept to each question and measure whether the model can still retain the correct answer under such perturbations. To ensure that distractors are not directly related, we select concepts drawn from different textbook chapters. For each question and each K, we sample one fixed distractor set once (single random seed) and use the same set for all models to enable paired comparisons.

To quantify retention under perturbation, we report **RUD**$_K$ (Retention Under Distractors): accuracy on perturbed items restricted to questions a model already solved without perturbation. Formally, letting S_m be the questions solved by model m in the unperturbed setting, $x_i^{(K)}$ be the question with K distractors prepended and y_i be the correct label,

$$\mathrm{RUD}_K(m) \;=\; \frac{1}{|S_m|} \sum_{i \in S_m} \mathbf{1}\Big\{ m\Big(x_i^{(K)}\Big) = y_i \Big\}.$$

We evaluate four models—Vanilla, CORE-Base, CORE-CR, and CORE-KL—under two splits: **Common** (items solved by all models; $n = 48$) and **Individual** (per-model solved sets: Vanilla 65 / CORE-Base 71 / CORE-CR 73 / CORE-KL 78). The resulting RUD_K curves for each split can be seen in Figure 2.

As K increases, models trained with CORE show consistently smaller accuracy drops than Vanilla and CORE-Base on both splits, with the CORE-CR variant particularly robust. This pattern indicates that CORE not only improves headline accuracy but also improves the robustness, making predictions more stable against irrelevant concept perturbations.

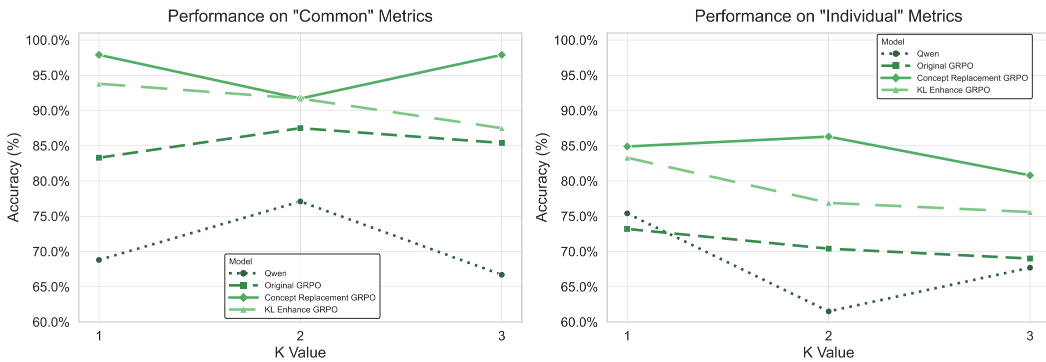

Figure 2: Performance comparison on Common vs Individual metrics.

5.3 DOES CORE APPLY TO BASE AND INSTRUCTION-TUNED MODELS?

We ask whether these gains persist across both base and instruction-tuned models. We therefore apply CORE to a base math model (Qwen2-Math-7B) and an instruction-tuned model (Llama-3-8B-Instruct), and evaluate under the same SC@21 (T=0.7) protocol across diverse out-of-domain suites (Tables 3). On Qwen2-Math-7B, comparing CORE-CR to Vanilla shows: GSM8K $+1.5$, ASDiv $+0.5$, MAWPS $+0.3$, TabMWP $+0.5$, MATH $+1.8$, MMLU-STEM -1.0, Gaokao -1.3, and OlympiadBench $+2.2$. On Llama-3-8B-Inst, CORE-CR improves GSM8k $+2.0$, ASDiv $+2.7$, MAWPS $+2.0$, and GaoKao $+3.2$, with small regressions on TabMWP (-3.7), MMLU-STEM (-1.3), and OlympiadBench (-3.7), and parity on MATH. Taken together, these results indicate that CORE is *model-agnostic*: it complements both base and instruct models, yielding good accuracy and stability across different domains.

6 CONCLUSION

In this work, we introduced **CORE (Concept-Oriented REinforcement)**, a reinforcement learning framework designed to bridge the definition–application gap in mathematical reasoning. By curating a high-quality concept–exercise corpus, diagnosing the limits of current LLMs, and injecting explicit concept signals into training via concept-aligned quizzes, concept-guided trajectory replacement, and KL-based divergence regularization, CORE provides fine-grained supervision beyond outcome correctness.

Extensive experiments on both base and instruction-tuned models demonstrate that CORE consistently improves performance on in-domain and out-of-domain benchmarks, yielding gains in concept selection, application, and robustness under perturbations. Importantly, these improvements arise without architectural modifications and are compatible with standard policy-gradient methods, underscoring the generality of the framework.

Our findings highlight that explicitly grounding reinforcement learning in mathematical concepts can substantially enhance the reasoning capabilities of LLMs, moving them beyond surface heuristics toward genuine conceptual competence. We hope this work motivates further exploration of concept-centered training signals, not only in mathematics but also across domains where principled reasoning is essential.

Ethics Statement

Our dataset is curated from high-quality educational resources originally published in Chinese. We contacted the primary author, who indicated they could not grant permission at this time due to unclear regulations around LLM training and evaluation in Copyright Law of China. After consulting legal guidance, we understand that limited use of such materials for non-commercial academic research may be permissible. Accordingly, our use is strictly for research and education; we do not redistribute substantial verbatim text. The research artifacts (codes, prompts, scripts, structured concept–exercise mappings, and model-generated quizzes/snippets) are derived and only small illustrative samples, that does not contain substantial portions of the original expression, would be presented. We cite sources and will promptly honor takedown or correction requests. Any released artifacts are for research use only and may not be used commercially; parties seeking commercial use should contact the rights holders. This statement is not legal advice, and we will adjust our practices as regulations evolve.

Reproducibility Statement

Our proposed framework is specified in §3. Our data curation, motivation verification, and training recipes are illustraed under it with subsections named as *Dataset Curation*, *Gap Diagnostics*, and *Concept Reinforcement Recipe*. The training and evaluation settings appear in §4 and Appendix B. We will provide a full code repo with codes, configs, and scripts to run training and evaluation end-to-end in the camera-ready version.

References

Anthropic. Claude 3.7 sonnet and claude code. Anthropic News, 2025. URL https://www.anthropic.com/news/claude-3-7-sonnet. Accessed: 2025-09-22.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics, 2024. URL https://arxiv.org/abs/2310.10631.

Luoxin Chen, Jinming Gu, Liankai Huang, Wenhao Huang, Zhicheng Jiang, Allan Jie, Xiaoran Jin, Xing Jin, Chenggang Li, Kaijing Ma, Cheng Ren, Jiawei Shen, Wenlei Shi, Tong Sun, He Sun, Jiahui Wang, Siran Wang, Zhihong Wang, Chenrui Wei, Shufa Wei, Yonghui Wu, Yuchen Wu, Yihang Xia, Huajian Xin, Fan Yang, Huaiyuan Ying, Hongyi Yuan, Zheng Yuan, Tianyang Zhan, Chi Zhang, Yue Zhang, Ge Zhang, Tianyun Zhao, Jianqiu Zhao, Yichi Zhou, and Thomas Hanwen Zhu. Seed-prover: Deep and broad reasoning for automated theorem proving, 2025. URL https://arxiv.org/abs/2507.23726.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. TheoremQA: A theorem-driven question answering dataset. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7889–7901, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.489. URL https://aclanthology.org/2023.emnlp-main.489/.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025a. URL https://arxiv.org/abs/2501.12948.

DeepSeek-AI. Deepseek-v3 technical report, 2025b. URL https://arxiv.org/abs/2412.19437.

Yu Feng, Ben Zhou, Weidong Lin, and Dan Roth. Bird: A trustworthy bayesian inference framework for large language models, 2025. URL https://arxiv.org/abs/2404.12494.

Team Gemini. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Dadi Guo, Jiayu Liu, Zhiyuan Fan, Zhitao He, Haoran Li, Yumeng Wang, and Yi R. Fung. Mathematical proof as a litmus test: Revealing failure modes of advanced large reasoning models, 2025a. URL https://arxiv.org/abs/2506.17114.

Jiaxing Guo, Wenjie Yang, Shengzhong Zhang, Tongshan Xu, Lun Du, Da Zheng, and Zengfeng Huang. Right is not enough: The pitfalls of outcome supervision in training llms for math reasoning, 2025b. URL https://arxiv.org/abs/2506.06877.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021a. URL https://arxiv.org/abs/2009.03300.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021b. URL https://openreview.net/forum?id=7Bywt2mQsCe.

Kaixuan Huang, Jiacheng Guo, Zihao Li, Xiang Ji, Jiawei Ge, Wenzhe Li, Yingqing Guo, Tianle Cai, Hui Yuan, Runzhe Wang, Yue Wu, Ming Yin, Shange Tang, Yangsibo Huang, Chi Jin, Xinyun Chen, Chiyuan Zhang, and Mengdi Wang. MATH-perturb: Benchmarking LLMs' math reasoning abilities against hard perturbations. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=OZy70UggXr.

Yichen Huang and Lin F. Yang. Gemini 2.5 pro capable of winning gold at imo 2025, 2025. URL https://arxiv.org/abs/2507.15855.

Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. MAWPS: A math word problem repository. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1152–1157, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1136. URL https://aclanthology.org/N16-1136/.

Yinghui Li, Jiayi Kuang, Haojing Huang, Zhikun Xu, Xinnian Liang, Yi Yu, Wenlian Lu, Yangning Li, Xiaoyu Tan, Chao Qu, Ying Shen, Hai-Tao Zheng, and Philip S. Yu. One example shown, many concepts known! counterexample-driven conceptual reasoning in mathematical LLMs. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=A31Ep22iQ7.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=DHyHRBwJUTN.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, Yansong Tang, and Dongmei Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct, 2025. URL https://arxiv.org/abs/2308.09583.

Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. A diverse corpus for evaluating and developing english math word problem solvers, 2021. URL https://arxiv.org/abs/2106.15772.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models, 2025. URL https://arxiv.org/abs/2410.05229.

OpenAI. Openai o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

OpenAI. Openai o3 and o4-mini system card. OpenAI, 2025. URL https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf. Accessed: 2025-09-22.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.168. URL https://aclanthology.org/2021.naacl-main.168/.

Tian Qin, Core Francisco Park, Mujin Kwun, Aaron Walsman, Eran Malach, Nikhil Anand, Hidenori Tanaka, and David Alvarez-Melis. Decomposing elements of problem solving: What "math" does rl teach?, 2025. URL https://arxiv.org/abs/2505.22756.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024. URL https://arxiv.org/abs/2409.19256.

Martin A Simon. Studying mathematics conceptual learning: Student learning through their mathematical activity. *North American Chapter of the International Group for the Psychology of Mathematics Education*, 2011. URL https://eric.ed.gov/?id=ED585966.

LLaMA Team. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Tianwen Wei, Jian Luan, Wei Liu, Shuang Dong, and Bin Wang. Cmath: Can your language model pass chinese elementary school math test?, 2023. URL https://arxiv.org/abs/2306.16636.

Mingqi Wu, Zhihao Zhang, Qiaole Dong, Zhiheng Xi, Jun Zhao, Senjie Jin, Xiaoran Fan, Yuhao Zhou, Huijie Lv, Ming Zhang, Yanwei Fu, Qin Liu, Songyang Zhang, and Qi Zhang. Reasoning or memorization? unreliable results of reinforcement learning due to data contamination, 2025. URL https://arxiv.org/abs/2507.10532.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024a. URL https://arxiv.org/abs/2407.10671.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement, 2024b. URL https://arxiv.org/abs/2409.12122.

Musheng Yao and Qihong Xie. *Advanced Algebra (Chinese 3rd Edition, Exercise)*. Fudan University Press, 2015. ISBN 9787309117769. URL https://www.google.com/books/edition/%E9%AB%98%E7%AD%89%E4%BB%A3%E6%95%B0/cyICkAEACAAJ?hl=en_US.

Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Liu, Ziyi Wang, Yudong Wang, Zijian Wu, Shuaibin Li, Fengzhe Zhou, Hongwei Liu, Songyang Zhang, Wenwei Zhang, Hang Yan, Xipeng Qiu, Jiayu Wang, Kai Chen, and Dahua Lin. Internlm-math: Open math large language models toward verifiable reasoning, 2024.

Xiaodong Yu, Ben Zhou, Hao Cheng, and Dan Roth. Reasonagain: Using extractable symbolic programs to evaluate mathematical reasoning, 2024. URL https://arxiv.org/abs/2410.19056.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MAmmoTH: Building math generalist models through hybrid instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=yLClGs770I.

Xuechen Zhang, Zijian Huang, Yingcong Li, Chenshun Ni, Jiasi Chen, and Samet Oymak. Bread: Branched rollouts from expert anchors bridge sft & rl for reasoning, 2025. URL https://arxiv.org/abs/2506.17211.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models, 2023.

Ben Zhou, Hongming Zhang, Sihao Chen, Dian Yu, Hongwei Wang, Baolin Peng, Dan Roth, and Dong Yu. Conceptual and unbiased reasoning in language models, 2024. URL https://arxiv.org/abs/2404.00205.

Ben Zhou, Sarthak Jain, Yi Zhang, Qiang Ning, Shuai Wang, Yassine Benajiba, and Dan Roth. Self-supervised analogical learning using language models, 2025. URL https://arxiv.org/abs/2502.00996.

A Details for textbook data

A.1 Data Curation Details

Our data curation followed a multi-stage pipeline to ensure high fidelity. We first employed an OCR tool[2] to digitize the textbook. The concept and exercise sections then underwent a manual verification stage, with any recognition errors corrected using GPT-4o. Subsequently, the entire Chinese corpus was translated into English via GPT-4o, followed by another round of human verification on the key sections to ensure accuracy.

B Experiment Setup

B.1 More Training Details

In the main experiments, we train models for 3 epochs on four H100 GPUs with GRPO by the *verl* (Sheng et al., 2024) framework. The policy is optimized using Adam with an actor learning rate of 1×10^{-6}, and a standard KL-divergence penalty with a fixed coefficient of 0.001 is applied against

[2] https://github.com/opendatalab/MinerU

the reference policy. During training, we set the sampling temperature to 0.7, with a batch size of 128 and a mini-batch size of 32. For each prompt, four responses are generated to conduct GRPO updates. Rewards follow a binary scheme (1 for correct, 0 for incorrect). The maximum prompt length is capped at 1024, and the maximum response length is also set to 1024.

B.2 DETAILS OF EVAULATION DATASETS

All evaluation datasets are sourced from the Qwen2.5-Math evaluation repository[3]. Their details are summarized below.

GRADE SCHOOL & MIDDLE SCHOOL LEVEL

- GSM8k: A dataset of approximately 8,500 high-quality, linguistically diverse elementary school math word problems. These problems require 2 to 8 steps of reasoning to solve and primarily involve basic arithmetic operations ($+$, $-$, $\times$, $\div$). The purpose of this dataset is to evaluate a model's ability to perform multi-step mathematical reasoning.
- ASDiv: An English math word problem dataset that is diverse in both linguistic patterns and problem types. It aims to comprehensively evaluate the true capabilities of math problem solvers, preventing models from achieving high scores merely by "memorizing" solutions to similar problems. Each problem is annotated with its problem type and grade level.
- MAWPS: A collection of several thousand English math word problems sourced from various online educational websites. Its goal is to provide an extensible repository of math problems for researchers to use and expand upon, covering a variety of basic arithmetic and algebraic problems.
- TabMWP: A large-scale dataset containing over 38,000 math word problems, distinguished by the inclusion of a table as context for each problem. To solve these, a model must be able to retrieve, integrate, and perform multi-step mathematical reasoning on information from both textual and tabular sources.

HIGH SCHOOL LEVEL

- MATH: A dataset created by Dan Hendrycks et al., containing 12,500 problems from American high school math competitions (such as AMC 10, AMC 12, AIME). The problems cover multiple subjects including algebra, geometry, number theory, and counting & probability. Each problem includes a detailed solution written by a human expert in LaTeX format. Its difficulty is significantly higher than elementary school problems, making it a key benchmark for advanced mathematical reasoning.
- MMLU-STEM: MMLU includes 57 different subjects, and MMLU-STEM refers to the subset of subjects related to STEM (Science, Technology, Engineering, and Mathematics), such as college-level math, physics, chemistry, and computer science.
- Gaokao2023En: This dataset is derived from China's "Gaokao" (National College Entrance Examination) mathematics papers. It typically involves translating Chinese math problems into English to test a large model's ability to solve difficult math problems from different cultural and educational backgrounds.

COLLEGE LEVEL & BEYOND

- CounterMath: A university-level mathematical benchmark designed to evaluate a model's conceptual reasoning by requiring it to prove or disprove statements by providing counterexamples. It focuses on advanced topics in Algebra, Topology, Real Analysis, and Functional Analysis.
- TheoremQA: A theorem-driven question answering dataset created to evaluate an AI model's ability to apply scientific theorems to solve challenging problems. It contains 800 questions covering over 350 theorems from Mathematics, Physics, EE&CS, and Finance.

[3] `https://github.com/QwenLM/Qwen2.5-Math`

Competition Level

- Olympiad Bench: A benchmark of extremely challenging, Olympiad-level scientific problems in both mathematics and physics. It is designed to push the boundaries of AGI research and often includes multimodal elements, requiring models to interpret diagrams and perform complex, creative reasoning.

B.3 More Evaluation Details

For evaluation, we adopt the SC@21 setting with a sampling temperature of 0.7. Specifically, 21 responses are generated, and after discarding cases where the answer cannot be extracted, the final prediction is determined by majority voting. In the event of a tie, one of the tied candidates is selected uniformly at random.

CORE-Base	**CORE-Base**	**CORE-KL**
Input: Let A, B be orthogonal matrices, k be a nonzero real number, and P an invertible matrix. Which of the following is true? A) $A + B$ is also an orthogonal matrix B) kA is also an orthogonal matrix C) AB is also an orthogonal matrix D) $P^{-1}AP$ is also an orthogonal matrix	**Input:** **Concepts:** Orthogonal and Unitary Transformations: Orthogonal and Unitary Matrices An n-dimensional real matrix A is called an orthogonal matrix if $A'A = AA' = I_n$. An n-dimensional complex matrix U is called a unitary matrix if $\overline{U}'U = U\overline{U}' = I_n$. **Exercise:** Let A, B be orthogonal matrices, k be a nonzero real number, and P an invertible matrix. Which of the following is true? A) $A + B$ is also an orthogonal matrix B) kA is also an orthogonal matrix C) AB is also an orthogonal matrix D) $P^{-1}AP$ is also an orthogonal matrix	**Input:** Let A, B be orthogonal matrices, k be a nonzero real number, and P an invertible matrix. Which of the following is true? A) $A + B$ is also an orthogonal matrix B) kA is also an orthogonal matrix C) AB is also an orthogonal matrix D) $P^{-1}AP$ is also an orthogonal matrix
Output: Answer: C | Response: B . . . B) kA is also an orthogonal matrix. This is true. **If A is an orthogonal matrix and k is a nonzero real number, then kA is also an orthogonal matrix. This is because multiplying an orthogonal matrix by a scalar does not change its orthogonality.** C) AB is also an orthogonal matrix.This is true. If A and B are orthogonal matrices, then their product AB is also an orthogonal matrix. This is because the product of two orthogonal matrices is orthogonal.. . . The final answer is boxed{B}.	**Output:** Answer: C | Response: B . . . B) kA is also an orthogonal matrix This is true. **If A is an orthogonal matrix, then kA is also an orthogonal matrix for any nonzero real number k.** C) AB is also an orthogonal matrix This is true. If A and B are orthogonal matrices, then AB is also an orthogonal matrix. Therefore, the correct answer is B) kA is also an orthogonal matrix.	**Output:** Answer: C | Response: C . . . B) kA is not necessarily an orthogonal matrix. **If k is a nonzero real number, then kA is not orthogonal unless $k = pm1$. This is because the columns of kA are not necessarily orthogonal unit vectors.** C) AB is an orthogonal matrix. The product of two orthogonal matrices is also orthogonal. This is because the columns of AB are linear combinations of the columns of A, which are orthogonal unit vectors. Therefore, the columns of AB are also orthogonal unit vectors. . . . The correct answer is C) AB is also an orthogonal matrix. The final answer is boxed{C}.

Table 5: This is a case study comparing the original GRPO model, a variant augmented with concept prompts, and the model after **CORE-KL** training. The example illustrates that the original model tends to fall into a fixed reasoning pattern when dealing with orthogonal matrices. Even when provided with detailed conceptual derivations, the model still prefers to rely on patterns learned during pretraining. In contrast, the model trained with **CORE-KL** is able to break out of this fixed paradigm and effectively apply the relevant theorems of orthogonal matrices.

C ANLAYSIS

Below is an exmaple for analysis.

D USE OF LLM

We have only used LLM for language polishing purposes in the paper writing. We do not use LLM for idea formalization, or to an extent that it could be regarded as a contributor.