

Silence the Judge: Reinforcement Learning with Self-Verifier via Latent Geometric Clustering

Nonghai Zhang^{1,2*} † Weitao Ma^{1,3*} † Zhanyu Ma¹ Jun Xu^{1‡}
Jiuchong Gao^{1‡} Jinghua Hao¹ Renqing He¹ Jingwen Xu¹
¹Meituan, ²Peking University, ³Harbin Institute of Technology

Abstract

Group Relative Policy Optimization (GRPO) significantly enhances the reasoning performance of Large Language Models (LLMs). However, this success heavily relies on expensive external verifiers or human rules. Such dependency not only leads to significant computational costs and training latency, but also yields sparse rewards that hinder optimization efficiency. To address these challenges, we propose **Latent-GRPO**, a framework that derives intrinsic rewards directly from latent space geometry. Crucially, our empirical analysis reveals a compelling geometric property: terminal token representations of correct reasoning trajectories form dense clusters with high intra-class similarity, whereas incorrect trajectories remain scattered as outliers. In light of this discovery, we introduce the **Iterative Robust Centroid Estimation (IRCE)** algorithm, which generates dense, continuous rewards by mitigating magnitude fluctuations via spherical projection and estimating a robust “truth centroid” through iterative aggregation. Experimental results on multiple datasets show that our method maintains model performance while achieving a training speedup of over 2× compared to baselines. Furthermore, extensive results demonstrate strong generalization ability and robustness.

1 Introduction

Large Language Models (LLMs) (Zhao et al., 2023) have achieved remarkable success in tackling complex reasoning tasks (Hendrycks et al., 2021b; Cobbe et al., 2021; Chen et al., 2021). To further enhance these capabilities, Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) has been established as the standard paradigm for model alignment. Specifically, this

*Work done during internship at Meituan.

†Equal contribution.

‡Corresponding authors.

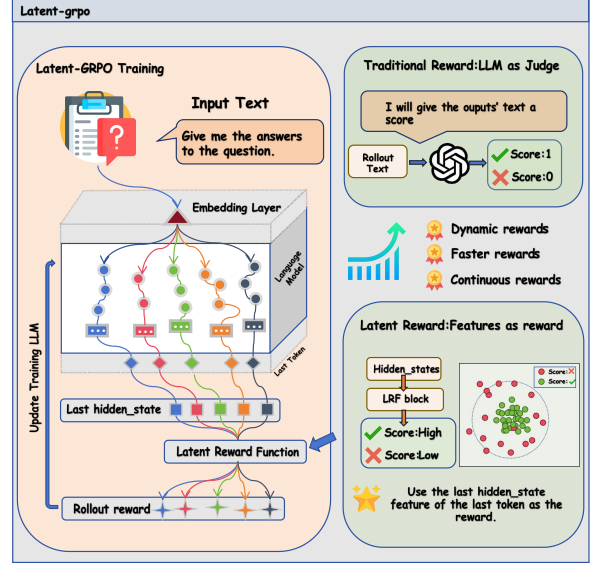


Figure 1: Comparison between conventional GRPO and Latent-GRPO. Conventional GRPO relies on expensive external verifiers to compute rewards, whereas Latent-GRPO autonomously extracts reward signals from the geometric structure of the latent space, eliminating external dependencies.

process has traditionally relied on Proximal Policy Optimization (PPO) (Rafailov et al., 2023) to refine policy performance through value-based reinforcement learning, while more recent approaches such as Group Relative Policy Optimization (GRPO) (Shao et al., 2024) further simplify the training pipeline by replacing the explicit value model with group-based relative advantages.

However, the practical efficacy of GRPO is often constrained by its heavy reliance on external verifiers (Wen et al., 2025; Zheng et al., 2023; Zhou et al., 2025), which makes training outcomes highly sensitive to their quality. On the one hand, rule-based verifiers are typically confined to deterministic tasks like mathematics. Moreover, designing clear and error-free rules for complex reasoning remains extremely difficult, and imperfect rules

can severely degrade training performance. On the other hand, employing external LLMs or training additional reward models incurs substantial costs (Lightman et al., 2023). These approaches introduce significant computational overhead and inference latency, which ultimately slows down the entire training process. Furthermore, such external judges are susceptible to biases or inaccurate scoring, thereby compromising training stability and final model quality (Xu et al., 2025; Cai et al., 2025). Beyond these concerns, most existing reward signals remain sparse and discrete (Tao et al., 2025). This binary feedback fails to capture the continuous semantic nuances of the reasoning process, often leading the model toward reward hacking (Cui et al., 2025; Gao et al., 2022). To address these challenges, we argue that an ideal reward mechanism should be **intrinsic, dense, and training-free**.

Building on representation engineering (Zou et al., 2023; Bartoszcze et al., 2025), which reveals that LLMs encode high-level semantic concepts internally (Marks and Tegmark, 2023), we discover a striking geometric property: the **last hidden states of terminal tokens** in correct reasoning trajectories form dense clusters, while incorrect paths remain scattered. Theoretically, this stems from the Transformer’s attention mechanism, which progressively aggregates the reasoning context into the final representation. Furthermore, this geometric consistency reflects the model’s inherent discriminative capabilities acquired during large-scale pre-training (Radford and Narasimhan, 2018). In essence, the latent space acts as an implicit verifier where logical consistency manifests as semantic convergence, providing a robust foundation for intrinsic reward modeling.

We introduce **Latent-GRPO**, a framework that utilizes robust intrinsic rewards from the geometric properties of the latent space, as shown in Figure 1. At its core, the **Iterative Robust Centroid Estimation (IRCE)** algorithm identifies a “truth centroid” from the **last hidden states** of terminal tokens, using their geometric relationship as a continuous reward. Unlike rule-based verifiers, our approach yields dense reward signals, which allows for more granular optimization across a wider range of reasoning scenarios where clear rules are unavailable. Compared to LLM-as-judge, Latent-GRPO eliminates external model dependencies, which significantly reduces training latency and prevents model collapse caused by inconsistent or noisy judging. More importantly, this approach

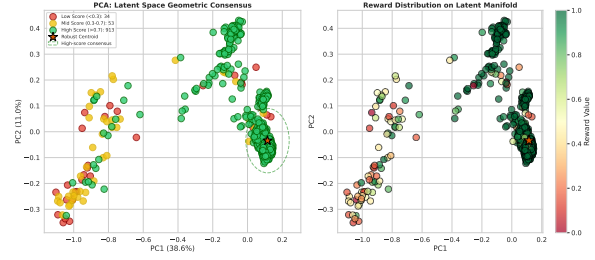


Figure 2: 2D PCA projection of 1,000 rollouts. Correct trajectories (green) form a dense consensus core around the truth centroid (gold star), while incorrect ones (red) scatter as outliers.

effectively activates the rich reasoning knowledge acquired by the model during its large-scale pre-training phase. Detailed experiments on GSM8K, MATH, and Open-Platypus all show that Latent-GRPO achieves 2× training speedup compared to LLM-as-Judge across three model scales (0.6B, 1.7B, and 4B). While it exceeds accuracy of both LLM-as-Judge and Rule-based Methods. Furthermore, additional analyses underscore the robustness and generalization ability of our approach across diverse scenarios.

2 Geometric Properties of Latent Space

This section demonstrates that the LLM latent space intrinsically captures reasoning quality through a series of empirical analyses. Specifically, we observe that correct reasoning trajectories exhibit high geometric clustering in their terminal hidden states, whereas incorrect paths remain scattered. Building on this, we explore leveraging these geometric features to score reasoning quality, ultimately finding a high degree of consistency with evaluations from external model-based verifiers. This alignment confirms the latent space as a robust and autonomous source of reward signals.

2.1 Theoretical Motivation

Our analysis is guided by two fundamental properties of the Transformer architecture. First, the last hidden state of the terminal token, h_T , acts as a semantic summary of the entire reasoning chain (Afzal et al., 2025; Wang et al., 2024c). Specifically, as the final representation before the language model head, h_T effectively aggregates the model’s converged reasoning information (Shai et al., 2024). Second, successful reasoning often leads to semantic collapse (Papayan et al., 2020; Wang et al., 2022). This means that while intermediate steps may vary, all correct trajectories tend

Framework of Latent-GRPO

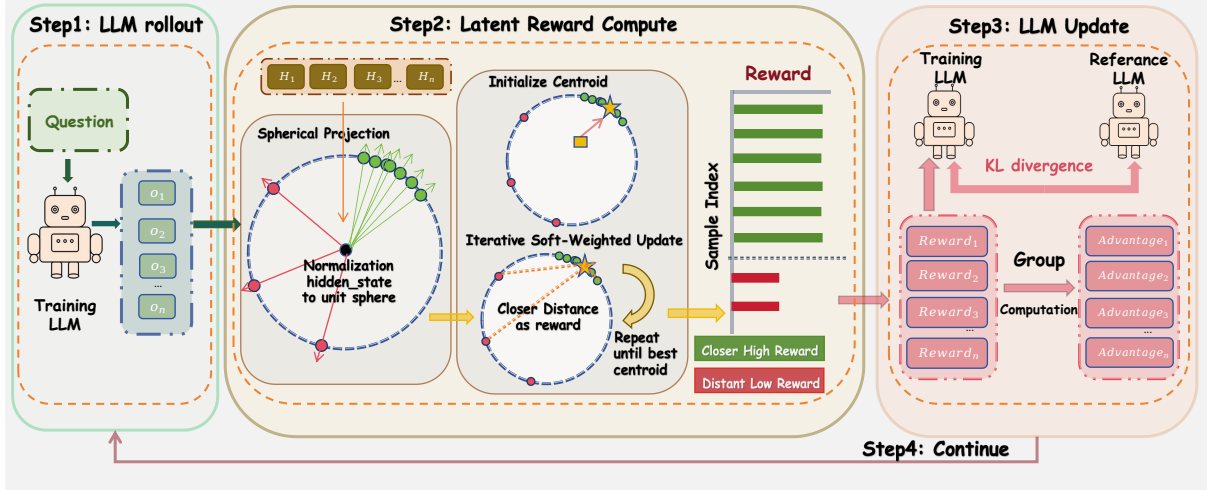


Figure 3: Overview of the Latent-GRPO framework. The policy model generates a group of responses for each prompt. Instead of relying on external verifiers, we extract the hidden states of the last token from each trajectory and apply the Iterative Robust Centroid Estimation (IRCE) algorithm to compute intrinsic rewards based on geometric clustering in the latent space. These rewards are then used to compute group-relative advantages for policy optimization. The entire process operates within the latent space, achieving zero additional inference overhead while providing dense reward signals.

to converge toward a unified semantic endpoint. In contrast, incorrect paths typically scatter due to their diverse failure modes (Marks and Tegmark, 2023; Zou et al., 2023).

2.2 Analysis Design

To validate our findings, we conduct two analyses on the GSM8K dataset. First, we generate 1,000 independent trajectories per prompt to examine global clustering patterns. Second, we construct GRPO groups ($G = 8$) to evaluate the effectiveness of geometric reward assignment. In both cases, we extract the last hidden state of the terminal token, $h_T \in \mathbb{R}^{1024}$, and compute distances directly in the original high-dimensional space to preserve the full semantic information. To verify these geometric rewards, we use GPT-4o to obtain ground-truth labels; critically, these labels are used only for validation in this section and are not required by our training framework.

2.3 Large-Scale Clustering Analysis

Figure 2 visualizes the **last hidden states** of terminal tokens via PCA, revealing a clear “core-periphery” structure.

Analyzing 1,000 trajectories from GSM8K, we observe three key patterns: (1) **High-Density Consensus Core**, where 913 correct trajectories cluster tightly around the centroid ($d_{\text{correct}} = 0.249$); (2)

Dispersed Outlier Region, where 53 incorrect trajectories remain scattered ($d_{\text{incorrect}} = 1.029$); and (3) 34 partially correct trajectories (score 0.3–0.9) fall between the two groups, demonstrating fine-grained geometric discrimination. The resulting $4.13 \times$ distance ratio confirms significant geometric separability of reasoning quality based on these **last hidden states**. To further demonstrate the universality of these patterns, we provide additional observations across model scales (from 0.6B to 4B) and diverse datasets (e.g., ScienceQA, ARB and Math) in Appendix I.1.

2.4 Validation via GRPO Group Simulation

Within each GRPO group, we calculate a robust centroid $\mathbf{c}$ via weighted aggregation of the terminal tokens’ **last hidden states**. The latent geometric score is then defined as $s_{\text{latent}}(i) = -\|\mathbf{h}_i - \mathbf{c}\|_2$. Figure 4 validates this approach: (a) presents a box plot of the reward distribution across different quality levels. It reveals that incorrect trajectories exhibit high dispersion and a lower median, whereas correct trajectories are significantly more concentrated with much higher reward values. (b-d) demonstrate high ranking consistency in representative groups. Notably, our method achieves a maximum Spearman correlation of $\rho = 0.927$ and a Top-1 selection agreement of up to 85% with external verifiers. These findings confirm that latent

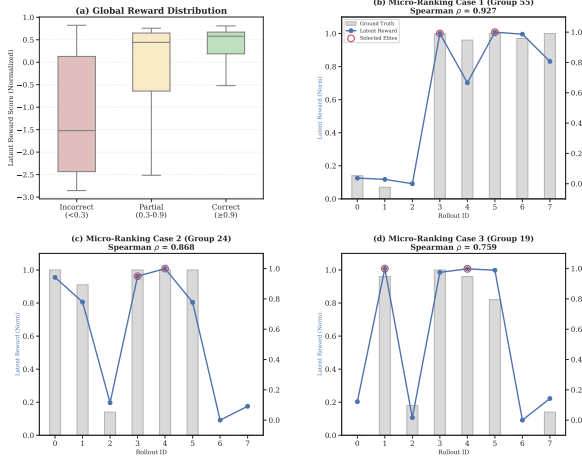


Figure 4: Validation of geometric-based scoring. (a) Distribution separability across quality levels (correct, partial, incorrect). (b-d) Group-level ranking consistency in representative 8-trajectory groups.

geometric properties can serve as a robust, training-free alternative to external judges, providing a solid foundation for Latent-GRPO. For a more detailed analysis, please refer to Appendix C. This property is further validated to persist throughout RL training in Section 5.

3 Methodology: Latent-GRPO

3.1 GRPO Short Review

Group Relative Policy Optimization (GRPO) (Shao et al., 2024) optimizes policy performance by computing relative advantages across a group of G trajectories $\{y_1, \dots, y_G\}$ sampled from the same prompt x :

$$A_i = \frac{R_i - \bar{R}_{\text{group}}}{\text{std}(R_{\text{group}}) + \epsilon} \quad (1)$$

where R_i represents the reward for the i -th trajectory. While architecturally efficient, standard GRPO faces two major challenges: heavy reliance on external verifiers, which leads to high costs and latency, and reward sparsity, which offers limited guidance for complex reasoning tasks. To address these issues, we propose a reward mechanism characterized by two key properties: (1) **intrinsic**, utilizing the model’s own **last hidden states of terminal tokens**, and (2) **dense**, providing continuous quality scores that enhance optimization gradients.

3.2 Core Algorithm: Iterative Robust Centroid Estimation

Based on the geometric findings in Section 2, we introduce the Iterative Robust Centroid Estimation (IRCE) algorithm to extract high-fidelity reward

signals. For each prompt, the policy model generates a group of G trajectories. We first extract the **last hidden states of terminal tokens** $\mathbf{h}_i \in \mathbb{R}^d$ ($i = 1, \dots, G$) and perform spherical normalization:

$$\tilde{\mathbf{h}}_i = \frac{\mathbf{h}_i}{\|\mathbf{h}_i\|_2} \quad (2)$$

This operation projects all trajectories onto a unit hypersphere, effectively eliminating magnitude fluctuations while ensuring the subsequent analysis focuses exclusively on semantic directionality.

The core of IRCE is to dynamically estimate a consensus centroid $\boldsymbol{\mu}$ that represents the “truth direction” of reasoning within the group. To suppress the influence of erroneous outliers, we employ an iterative soft-weighting mechanism. In each iteration s , we compute soft weights $w_i^{(s)}$ using a Gaussian kernel based on the sample’s current distance to the centroid:

$$w_i^{(s)} = \frac{\exp\left(-\left(d_i^{(s)}\right)^2 / 2\left(\sigma^{(s)}\right)^2\right)}{\sum_{j=1}^G \exp\left(-\left(d_j^{(s)}\right)^2 / 2\left(\sigma^{(s)}\right)^2\right)} \quad (3)$$

where $\sigma^{(s)}$ is an adaptive scale parameter derived from the group’s distance distribution. The centroid is subsequently updated via weighted aggregation: $\boldsymbol{\mu}^{(s+1)} = \text{Norm}\left(\sum w_i^{(s)} \tilde{\mathbf{h}}_i\right)$. **Upon convergence** after T iterations, the intrinsic reward R_i is defined as the negative Euclidean distance to the final centroid:

$$R_i = \frac{-\|\tilde{\mathbf{h}}_i - \boldsymbol{\mu}^{(T)}\|_2 - \min_j (-\|\tilde{\mathbf{h}}_j - \boldsymbol{\mu}^{(T)}\|_2)}{\max_j (-\|\tilde{\mathbf{h}}_j - \boldsymbol{\mu}^{(T)}\|_2) - \min_j (-\|\tilde{\mathbf{h}}_j - \boldsymbol{\mu}^{(T)}\|_2)} \quad (4)$$

This mechanism yields intrinsic rewards that are **dense, continuous, and naturally bounded**. By applying Min-Max normalization, we map the distances into a calibrated range $R_i \in [0, 1]$, which prevents gradient explosion and ensures optimization stability across different reasoning tasks. The complete procedure, including the adaptive scaling and normalization, is summarized in Algorithm 1. Furthermore, we provide a rigorous mathematical derivation of this iterative weighting scheme in Appendix D to demonstrate its robustness against sampling noise.

3.3 Latent-GRPO Framework

As illustrated in Figure 3, **Latent-GRPO** integrates the IRCE algorithm into the GRPO pipeline. Unlike traditional reinforcement learning which relies

Algorithm 1 Iterative Robust Centroid Estimation

Input: Hidden states $\{\mathbf{h}_1, \dots, \mathbf{h}_G\}$, max iterations T
Output: Rewards $\{R_1, \dots, R_G\}$

```
1: // Step 1: Spherical Projection
2: for  $i = 1$  to  $G$  do
3:    $\tilde{\mathbf{h}}_i \leftarrow \mathbf{h}_i / \|\mathbf{h}_i\|_2$ 
4: end for
5: // Step 2: Initialize Centroid
6:  $\boldsymbol{\mu} \leftarrow \frac{1}{G} \sum_{i=1}^G \tilde{\mathbf{h}}_i$ 
7:  $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} / \|\boldsymbol{\mu}\|_2$ 
8: // Step 3: Iterative Soft-Weighted Update
9: for  $s = 0$  to  $T - 1$  do
10:  for  $i = 1$  to  $G$  do
11:     $d_i \leftarrow \|\tilde{\mathbf{h}}_i - \boldsymbol{\mu}\|_2$ 
12:  end for
13:   $\sigma \leftarrow \text{std}(\{d_1, \dots, d_G\}) + \epsilon$ 
14:  for  $i = 1$  to  $G$  do
15:     $w_i \leftarrow \exp(-(d_i)^2 / (2\sigma^2))$ 
16:  end for
17:   $\mathbf{w} \leftarrow \mathbf{w} / \sum_j w_j$  ▷ Normalize weights
18:   $\boldsymbol{\mu} \leftarrow \sum_i w_i \tilde{\mathbf{h}}_i$ 
19:   $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} / \|\boldsymbol{\mu}\|_2$  ▷ Normalize centroid
20: end for
21: // Step 4: Compute Rewards
22: for  $i = 1$  to  $G$  do
23:    $d_i \leftarrow \|\tilde{\mathbf{h}}_i - \boldsymbol{\mu}\|_2$ 
24:    $R_i \leftarrow -d_i$ 
25: end for
26:  $\mathbf{R} \leftarrow \text{MinMaxNormalize}(\mathbf{R})$ 
27: return  $\mathbf{R}$ 
```

on static external verifiers, Latent-GRPO adopts a dynamic, intrinsic reward mechanism that operates without external supervision.

Dynamic Adaptability. As the policy model evolves during training, its latent representation space naturally shifts. Unlike static external verifiers, IRCE dynamically estimates the consensus centroid for each training batch, allowing reward signals to adapt in real-time to the model’s current state. This online adaptation effectively mitigates distribution shift—a common challenge when reward signals become misaligned with evolving model representations. Additionally, the iterative soft-weighting mechanism ensures stable and robust gradient signals even when initial sample quality is suboptimal.

Computational Efficiency. Latent-GRPO eliminates external verification overhead by operating on hidden states already computed during rollout. Unlike conventional verifiers requiring $O(GL)$ additional forward passes, IRCE runs with $O(GTd)$ complexity; since $T \ll L$ (typically 5 iterations vs. 2048 tokens), the overhead is negligible. This reduces memory footprint and transforms reward computation into an efficient intrinsic process.

4 Experimental Setup

4.1 Datasets and Evaluation Metrics

Datasets. We evaluate our approach on three complementary datasets that span different reasoning complexity levels and domains. **GSM8K** (Cobbe et al., 2021) contains elementary-level mathematical word problems, establishing a foundation for basic arithmetic reasoning. **MATH** (Hendrycks et al., 2021b) comprises high school and competition-level problems, testing more sophisticated mathematical reasoning. **Open-Platypus** (Lee et al., 2023a) covers diverse reasoning tasks across physics, logic, and mathematics, enabling validation across multiple domains. Together, these datasets provide comprehensive coverage for evaluating both accuracy improvements and computational efficiency across varying reasoning complexities. The split of train/test datasets and detailed dataset statistics are provided in Appendix E.

Evaluation Metrics. We measure two primary performance dimensions: (1) **Task Accuracy** on each test dataset of benchmark to assess reasoning capability, and (2) **Training Efficiency** measured as time per epoch across different reward methods under identical experimental settings. For evaluation, GSM8K uses rule-based answer matching; all other datasets use GPT-4o as the evaluation judge. This protocol is applied uniformly across all compared methods, ensuring a fair comparison.

4.2 Experimental Setup

Models and Hardware. We validate our method across three Qwen models of varying sizes: Qwen3-0.6B, Qwen3-1.7B, and Qwen3-4B. This range enables assessment of our method’s effectiveness across different model scales. We further provide additional experiments on larger-scale models (Qwen3-8B and Qwen3-14B) in Appendix I.2. For detailed hyperparameter settings including GRPO training configuration, IRCE algorithm parameters, and hardware specifications, refer to Appendix F.

Main Experiment We compare three reward paradigms across three datasets (GSM8K, MATH, Open-Platypus) and three model sizes (Qwen3-0.6B, Qwen3-1.7B, Qwen3-4B): (1) **LLM-as-Judge** using external verification (GPT-4o), (2) **Rule-based** verification with ground-truth labels, and (3) our proposed **Latent-GRPO** leveraging intrinsic latent space geometry. This comparison evaluates whether intrinsic geometric signals can match or exceed external verification in both ac-

Dataset	Method	Qwen3-0.6B		Qwen3-1.7B		Qwen3-4B	
		Acc $\uparrow$	Time $\downarrow$	Acc $\uparrow$	Time $\downarrow$	Acc $\uparrow$	Time $\downarrow$
<i>GSM8K</i>	LLM-as-Judge	53.52%	768.42m	64.20%	1032.55m	72.12%	1411.72m
	Rule-based	58.41%	434.61m	71.55%	488.63m	79.87%	651.45m
	Latent-GRPO (Ours)	61.25%	431.18m	73.88%	492.34m	82.34%	658.21m
<i>MATH</i>	LLM-as-Judge	52.94%	1224.15m	65.77%	1608.34m	77.44%	2357.31m
	Rule-based	55.63%	723.12m	42.14%	814.22m	62.63%	1084.72m
	Latent-GRPO (Ours)	58.47%	718.63m	78.51%	811.51m	77.53%	1081.47m
<i>Open-Platypus</i>	LLM-as-Judge	34.45%	1937.82m	56.69%	2573.41m	65.21%	3522.18m
	Latent-GRPO (Ours)	40.56%	1079.27m	64.82%	1218.92m	78.06%	1632.52m

Table 1: Comprehensive comparison of reward methods across datasets and model scales. The results demonstrate that Latent-GRPO consistently achieves superior accuracy and training efficiency (Time per epoch) compared to LLM-as-Judge and Rule-based baselines. (Simulated QPS for LLM-as-Judge is 2)

Method	Acc $\uparrow$	Time (m) $\downarrow$
<i>Qwen3-0.6B</i>		
Mean Pooling	58.74%	435.22
Weighted Mean	57.12%	442.89
Last Token (Ours)	61.25%	431.18
<i>Qwen3-1.7B</i>		
Mean Pooling	71.05%	497.61
Weighted Mean	69.88%	512.14
Last Token (Ours)	73.88%	492.34
<i>Qwen3-4B</i>		
Mean Pooling	79.45%	664.33
Weighted Mean	78.12%	685.56
Last Token (Ours)	82.34%	658.21

Table 2: Comparison of different hidden state extraction methods. Using the terminal token’s representation (Last Token) consistently yields the best reasoning performance with the lowest computational latency.

curacy and training efficiency. Detailed baseline descriptions are provided in Appendix G.1.

5 Results and Analysis

5.1 Reward Methods Comparison

We compare three reward paradigms on GSM8K and MATH: LLM-as-Judge (GPT-4o), Rule-based verification, and Latent-GRPO. For Open-Platypus, we focus on LLM-as-Judge and Latent-GRPO since rule-based methods are limited to tasks with deterministic ground truth.

As shown in Table 1, Latent-GRPO achieves approximately $2\times$ **training speedup** compared to LLM-as-Judge across all datasets and model scales. Meanwhile, Latent-GRPO maintains or exceeds the accuracy of both LLM-as-Judge and Rule-based verification. Detailed per-dataset and per-model results are provided in Appendix H.1.

Efficiency. Latent-GRPO eliminates the external verifier bottleneck that constrains LLM-as-Judge.

The latter faces two system-level costs: API rate limiting (2 QPS) forces scoring requests into a queue, and each API call incurs 1-2 minutes of latency. Together, these consume 58-63% of total training time. Latent-GRPO instead computes rewards on hidden states already available from the forward pass using the IRCE algorithm, requiring only geometric operations with $O(Td)$ complexity (where $T \ll L$, typically 5 iterations vs. 2048 sequence length), compared to $O(GL)$ for transformer-based verifiers. Rule-based verification operates at similar cost to Latent-GRPO, confirming that speedup comes primarily from eliminating external calls. Fairness comparisons against higher-QPS LLM-as-Judge and a trained reward model (ArmoRM(Wang et al., 2024b)), alongside multi-run statistical significance tests, are provided in Appendix G.1 and Appendix K. Latent-GRPO remains faster even at QPS=10, and all accuracy improvements are statistically significant ($p < 0.05$). **Accuracy.** This advantage stems from two factors: Latent-GRPO’s continuous, centroid-based rewards provide richer gradient signals than the binary 0/1 feedback of Rule-based and LLM-as-Judge methods, while deriving rewards from the model’s own geometry avoids external inconsistency and noise, leading to more stable training and reducing the risk of training collapse.

5.2 Ablation Studies: Core Design Choices

We validate two critical design choices through systematic ablation studies: (1) The hidden state extraction method for capturing reasoning quality, and (2) The centroid estimation algorithm for computing intrinsic rewards. Detailed quantitative analysis and per-model comparisons are provided in Appendix H.2.

Method	Acc $\uparrow$	Time (m) $\downarrow$
Qwen3-0.6B		
Mean Pool	57.12%	452.34
K-Means	58.85%	489.12
Eigen Centrality	59.43%	468.76
IRCE (Ours)	61.25%	431.18
Qwen3-1.7B		
Mean Pool	68.45%	512.67
K-Means	70.12%	543.89
Eigen Centrality	71.56%	531.42
IRCE (Ours)	73.88%	492.34
Qwen3-4B		
Mean Pool	77.89%	682.12
K-Means	79.23%	725.67
Eigen Centrality	80.56%	708.45
IRCE (Ours)	82.34%	658.21

Table 3: Comparison of different latent consensus scoring methods on GSM8K. IRCE (Ours) achieves the highest reasoning accuracy with the minimum computational overhead across all model scales.

Hidden State Extraction. We compare three approaches for extracting reasoning quality from hidden states: (1) **Last Token**, using only the terminal token’s representation; (2) **Mean Pooling**, averaging representations across all tokens; (3) **Weighted Mean**, computing weighted averages of key tokens (weighted averaging of pre-designed key tokens such as mathematical operators, structural markers, and reasoning keywords).

As illustrated in Table 2, the Last Token method consistently outperforms all aggregation baselines across various model scales. This finding reveals a fundamental property of transformer-based reasoning: the final token, as the immediate precursor to the end-of-sequence (EOS) prediction, acts as a semantic bottleneck where reasoning correctness is crystallized. In contrast, Mean Pooling degrades performance by incorporating noise from intermediate tokens that are often orthogonal to the final correctness. The failure of Weighted Mean, which focuses on linguistically significant keywords (e.g., "therefore", "solution") is particularly instructive. It suggests that reasoning quality is not localized at specific lexical markers but emerges holistically through the generation process, ultimately converging into the final representation.

Centroid Estimation We compare four approaches for estimating the consensus centroid from a group of trajectories: (1) **Mean Pooling**, (2) **K-Means clustering**, (3) **Eigen Centrality**, and (4) **IRCE**, our proposed iterative robust centroid estimation. All experiments are conducted on GSM8K across

three model scales using the Last Token extraction method identified above. For details of these latent reward methods, please refer to Appendix G.2.

As shown in Table 3, IRCE consistently outperforms all baseline methods across all model scales. Mean Pooling fails to handle outliers effectively, resulting in suboptimal centroids. K-Means attempts to improve through hard cluster assignment and achieves competitive accuracy, yet still underperforms IRCE. Eigen Centrality uses graph-based importance weighting but introduces significant computational overhead through eigendecomposition, achieving lower accuracy while incurring higher computational cost. Our method, IRCE, maintains both robustness and efficiency, making it the optimal choice for latent reward estimation. Sensitivity analyses for group size G and iteration count T are provided in Appendix H.2, results remain stable across $G \in \{6, 8, 10\}$ and $T \in \{3, 5, 7\}$.

5.3 Generalization

Capability Preservation. A critical concern in RL training is the potential loss of general capabilities due to task-specific overfitting. To address this, we train Latent-GRPO on a specific mixture of reasoning datasets and subsequently evaluate it on a suite of unseen general benchmarks, including MMLU, AIME (24 & 25), BBH, and MATH-500. As illustrated in Table 4, Latent-GRPO consistently maintains or surpasses the performance of base models across these diverse tasks. These results confirm that dense intrinsic rewards guide the model toward mastering transferable reasoning patterns rather than overfitting. For analysis of the mixed AIME results, see Appendix H.1.

Cross-Model Generalization. To validate that Latent-GRPO generalizes beyond the Qwen3 series, we extend our evaluation to Llama3.2-3B on the GSM8K, MATH, and Open-Platypus datasets. As shown in Table 5, the latent space geometry provides a universal signal for reasoning quality that is independent of the model family.

Scaling to Larger Models. To validate behavior at greater scale, we further evaluate Latent-GRPO on Qwen3-8B and Qwen3-14B (Appendix I.2). Latent-GRPO achieves pass@1 of 0.77 and 0.81 respectively, with training time approximately 40–50% of LLM-as-Judge, confirming that geometric clustering and efficiency gains generalize well beyond the 4B scale.

Method	AIME24 $\uparrow$	AIME25 $\uparrow$	MATH-500 $\uparrow$	MMLU $\uparrow$	BBH $\uparrow$	Avg $\uparrow$	Time (m) $\downarrow$
Qwen3-0.6B							
Base	10.7	15.1	77.6	52.8	41.5	39.54	—
GRPO (LLM-Judge)	9.4	11.7	65.9	50.6	42.4	36.00	4082.96
Latent-GRPO (Ours)	10.6	19.2	67.3	49.9	39.2	37.24	2280.52
Qwen3-1.7B							
Base	48.3	36.8	93.4	62.6	54.5	59.12	—
GRPO (LLM-Judge)	48.1	33.8	90.6	67.6	67.4	61.50	4988.84
Latent-GRPO (Ours)	44.6	35.4	91.2	68.1	68.7	61.60	2340.05
Qwen3-4B							
Base	73.8	65.6	97.0	83.7	72.6	78.54	—
GRPO (LLM-Judge)	72.9	63.3	96.4	85.5	81.9	80.00	6753.47
Latent-GRPO (Ours)	74.6	66.7	97.5	88.5	82.3	81.92	3108.44

Table 4: Performance and efficiency comparison across Qwen3 scales on mathematics (AIME, MATH-500) and general reasoning (MMLU, BBH) benchmarks.

Method	Acc $\uparrow$	Time (m) $\downarrow$
GSM8K (2k steps)		
LLM-as-Judge	71.34%	1284.56
Latent-GRPO	78.62%	591.24
MATH (3k steps)		
LLM-as-Judge	43.12%	1945.12
Latent-GRPO	52.45%	912.87
Open-Platypus (4.5k steps)		
LLM-as-Judge	61.88%	2915.68
Latent-GRPO	73.12%	1386.42

Table 5: Experimental results of Latent-GRPO on Llama3.2-3B. Following the same protocol as Qwen3 experiments, Latent-GRPO demonstrates superior scaling performance and significantly reduced training latency compared to the LLM-as-judge baseline.

5.4 Further Analysis

To verify that the clustering assumption holds throughout actual RL training, we randomly sampled 200 rollout groups from training logs across three model scales on GSM8K, MATH, and Open-Platypus, with correctness annotated by GPT-4o-mini. As shown in Table 6, correct trajectories consistently maintain low centroid distances (0.23–0.29), while incorrect trajectories remain substantially farther away (1.01–1.13), resulting in stable distance ratios of 3.9–4.4 $\times$. In addition, each rollout group contains an average of 4.3–5.6 correct responses, consistently exceeding the majority threshold required for reliable centroid estimation. These results indicate that the risk of reward misalignment, where an incorrect consensus dominates the reward signal, remains empirically negligible throughout training. Further analysis of intermediate step quality is provided in Appendix L.

Furthermore, Latent-GRPO reduces output di-

Dataset	Scale	Avg $\checkmark$	d_{cor}	d_{inc}	Ratio
GSM8K	0.6B	4.72	0.256	1.012	3.95 $\times$
	1.7B	5.18	0.242	1.045	4.32 $\times$
	4B	5.56	0.238	1.056	4.44 $\times$
MATH	0.6B	4.35	0.289	1.123	3.89 $\times$
	1.7B	4.87	0.267	1.089	4.08 $\times$
	4B	5.23	0.254	1.078	4.24 $\times$
Open-Platypus	0.6B	4.87	0.283	1.112	3.93 $\times$
	1.7B	5.23	0.269	1.087	4.04 $\times$
	4B	5.58	0.258	1.065	4.13 $\times$

Table 6: In-training clustering statistics from 200 sampled rollout groups per setting. “Avg $\checkmark$ ” denotes average correct count per group; $d_{\text{cor}}/d_{\text{inc}}$: average centroid distance of correct/incorrect trajectories.

versity by only 3.0–4.2%, compared to 17.6–20.0% for LLM-as-Judge, demonstrating that the proposed dense reward mechanism effectively mitigates mode collapse (Appendix J).

Additional qualitative case studies comparing reward signals are provided in Appendix M.

6 Conclusion

In this work, we introduce Latent-GRPO, a training framework designed to overcome the efficiency bottlenecks and sparse rewards inherent in verifier-dependent reinforcement learning. At the core of this framework lies the Iterative Robust Centroid Estimation algorithm, which transforms latent geometry into dense rewards, eliminating the reliance on slow external verifiers and achieving a substantial acceleration in training compared to the LLM-as-a-judge baseline. Furthermore, extensive evaluations on unseen benchmarks confirm that our approach maintains competitive accuracy. It effectively preserves native capabilities and prevents

task-specific overfitting. These findings substantiate that LLMs possess inherent self-evaluation mechanisms and offer a scalable paradigm for verifier-free post-training. A comprehensive discussion of related work is provided in Appendix B.

Limitations

Our work has two primary limitations that we aim to address in future research. First, while **Latent-GRPO** is effective up to 14B parameters, its scaling behavior in ultra-large models (70B+) and applicability to open-ended generation remain to be explored. Second, while we provide strong empirical evidence, a formal mathematical framework for latent clustering is still nascent. Moving forward, we plan to extend the geometric consensus hypothesis to broader tasks and investigate hybrid frameworks that integrate **IRCE** rewards with offline paradigms like DPO to further stabilize self-supervised alignment.

References

- Anum Afzal, Florian Matthes, Gal Chechik, and Yftah Ziser. 2025. [Knowing before saying: Llm representations encode information about chain-of-thought success before completion](#). *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12791–12806.
- Lukasz Bartoszcze, Sarthak Munshi, Bryan Sukidi, Jennifer Yen, Zejia Yang, David Williams-King, Linh Le, Kosi Asuzu, and Carsten Maple. 2025. [Representation engineering for large-language models: Survey and research challenges](#). *arXiv.org*.
- Zheni Bi, Kai Han, Chuanjian Liu, Yehui Tang, and Yunhe Wang. 2024. [Forest-of-thought: Scaling test-time compute for enhancing llm reasoning](#). *International Conference on Machine Learning*.
- Xin-Qiang Cai, Wei Wang, Feng Liu, Tongliang Liu, Gang Niu, and Masashi Sugiyama. 2025. [Reinforcement learning with verifiable yet noisy rewards under imperfect verifiers](#). *arXiv.org*.
- Haolin Chen, Yihao Feng, Zuxin Liu, Weiran Yao, Akshara Prabhakar, Shelby Heinecke, Ricky Ho, P. Mui, Silvio Savarese, Caiming Xiong, and 1 others. 2024. [Language models are hidden reasoners: Unlocking latent reasoning capabilities via self-rewarding](#). *arXiv.org*.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé, Jared Kaplan, Harrison Edwards, Yura Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. Evaluating large language models trained on code. *arXiv.org*.
- K. Cobbe, Vineet Kosaraju, Mo Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv.org*.
- Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, and 1 others. 2025. [Process reinforcement through implicit rewards](#). *Trans. Mach. Learn. Res.*
- Hanwen Du, Yuxin Dong, and Xia Ning. 2025. [Latent thinking optimization: Your latent reasoning language model secretly encodes reward signals in its latent thoughts](#). *arXiv.org*.
- Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, M. Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, and 1 others. 2021. Glam: Efficient scaling of language models with mixture-of-experts. In *International conference on machine learning*, pages 5547–5569. PMLR.
- Leo Gao, John Schulman, and Jacob Hilton. 2022. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR.
- Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. 2025. [Scaling up test-time compute with latent reasoning: A recurrent depth approach](#). *Advances in Neural Information Processing Systems 38*, pages 46242–46293.
- Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. 2023. [Think before you speak: Training language models with pause tokens](#). *International Conference on Learning Representations*.
- Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. [Training large language models to reason in a continuous latent space](#). *arXiv.org*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, D. Song, and J. Steinhardt. 2020. Measuring massive multitask language understanding. *International Conference on Learning Representations*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, D. Song, and J. Steinhardt. 2021a. Measuring mathematical problem solving with the math dataset. *NeurIPS Datasets and Benchmarks*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *NeurIPS Datasets and Benchmarks*.

- Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. [Adaptive mixtures of local experts](#). *Neural computation*, 3(1):79–87.
- Dhiraj D. Kalamkar, Dheevatsa Mudigere, Naveen Mellempudi, Dipankar Das, K. Banerjee, Sasikanth Avancha, Dharma Teja Vooturi, Nataraj Jammalamadaka, Jianyu Huang, Hector Yuen, and 1 others. 2019. A study of bfloat16 for deep learning training. *arXiv.org*.
- Ariel N Lee, Cole J Hunter, and Nataniel Ruiz. 2023a. [Platypus: Quick, cheap, and powerful refinement of llms](#). *arXiv.org*.
- Harrison Lee, Samrat Phatale, H. Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023b. [Rlaif: Scaling reinforcement learning from human feedback with ai feedback](#). *arXiv.org*.
- Yi-Chen Li, Tian Xu, Yang Yu, Xuqin Zhang, Xiong-Hui Chen, Zhongxiang Ling, Ningjing Chao, Lei Yuan, and Zhi-Hua Zhou. 2025. [Generalist reward models: Found inside large language models](#). *arXiv.org*.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. [Let’s verify step by step](#). In *International Conference on Learning Representations*.
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. [Understanding rl-zero-like training: A critical perspective](#). *arXiv.org*.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. [Self-refine: Iterative refinement with self-feedback](#). *Advances in Neural Information Processing Systems* 36, 36:46534–46594.
- Samuel Marks and Max Tegmark. 2023. [The geometry of truth: Emergent linear structure in large language model representations of true/false datasets](#). *arXiv.org*.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, S. Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. [Training language models to follow instructions with human feedback](#). *Neural Information Processing Systems*, 35:27730–27744.
- Vardan Papyan, XY Han, and David L Donoho. 2020. [Prevalence of neural collapse during the terminal phase of deep learning training](#). *Proceedings of the National Academy of Sciences of the United States of America*, 117(40):24652–24663.
- Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). *Neural Information Processing Systems*, 36:53728–53741.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv.org*.
- Adam Shai, Lucas Teixeira, Alexander Oldenziel, Sarah Marzen, and Paul Riechers. 2024. [Transformers represent belief state geometry in their residual stream](#). *Advances in Neural Information Processing Systems* 37, 37:75012–75034.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, R. Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *arXiv.org*.
- Nisan Stiennon, O. Long, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and P. Christiano. 2020. Learning to summarize with human feedback. *Neural Information Processing Systems*, 33:3008–3021.
- Hanshi Sun, Momin Haider, Ruiqi Zhang, Huitao Yang, Jiahao Qiu, Ming Yin, Mengdi Wang, Peter Bartlett, and Andrea Zanette. 2024. [Fast best-of-n decoding via speculative rejection](#). *Advances in Neural Information Processing Systems* 37, 37:32630–32652.
- Mirac Suzgun, Nathan Scales, Nathanael Scharli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, A. Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and 1 others. 2023. [Challenging big-bench tasks and whether chain-of-thought can solve them](#). *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051.
- Leitian Tao, Ilia Kulikov, Swarnadeep Saha, Tianlu Wang, Jing Xu, Sharon Li, Jason E Weston, and Ping Yu. 2025. [Hybrid reinforcement: When reward is sparse, it’s better to be dense](#). *arXiv.org*.
- Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. 2024a. [Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8642–8655. Association for Computational Linguistics.
- Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024b. [Interpretable preferences via multi-objective reward modeling and mixture-of-experts](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10582–10592. Association for Computational Linguistics.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, and Denny Zhou. 2022. Self-consistency

- improves chain of thought reasoning in language models. *International Conference on Learning Representations*.
- Yiming Wang, Pei Zhang, Baosong Yang, Derek F Wong, and Rui Wang. 2024c. [Latent space chain-of-embedding enables output-free llm self-evaluation](#). *International Conference on Learning Representations*.
- Xumeng Wen, Zihan Liu, Shun Zheng, Zhijian Xu, Shengyu Ye, Zhirong Wu, Xiao Liang, Yang Wang, Junjie Li, Ziming Miao, and 1 others. 2025. [Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms](#). *arXiv.org*.
- Zhangchen Xu, Yuetai Li, Fengqing Jiang, Bhaskar Ramasubramanian, Luyao Niu, Bill Yuchen Lin, and Radha Poovendran. 2025. [Tinyv: Reducing false negatives in verification improves rl for llm reasoning](#). *arXiv.org*.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. [Tree of thoughts: Deliberate problem solving with large language models](#). *Advances in Neural Information Processing Systems* 36, 36:11809–11822.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, and 1 others. 2025. [Dapo: An open-source llm reinforcement learning system at scale](#). *Advances in Neural Information Processing Systems* 38, pages 125532–125554.
- Yizhe Zhang, Jiatao Gu, Zhuofeng Wu, Shuangfei Zhai, Joshua Susskind, and Navdeep Jaitly. 2023. [Planner: Generating diversified paragraph via latent language diffusion model](#). *Advances in Neural Information Processing Systems* 36, 36:80178–80190.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. [A survey of large language models](#). *Frontiers Comput. Sci.*, 1(2).
- Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Y. Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, and 1 others. 2025. [Geometric-mean policy optimization](#). *arXiv.org*.
- Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, K. Dang, Yuqiong Liu, Rui Men, An Yang, and 1 others. 2025. [Group sequence policy optimization](#). *arXiv.org*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, E. Xing, and 1 others. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). *Neural Information Processing Systems*, 36:46595–46623.
- Xiangxin Zhou, Zichen Liu, Anya Sims, Haonan Wang, Tianyu Pang, Chongxuan Li, Liang Wang, Min Lin, and Chao Du. 2025. [Reinforcing general reasoning without verifiers](#). *arXiv.org*.
- Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv.org*.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. [Representation engineering: A top-down approach to ai transparency](#). *arXiv.org*.

A Usage of LLM

Large Language Models were used exclusively to improve the clarity and fluency of English writing. They were not involved in research ideation, experimental design, data analysis, or interpretation. The authors take full responsibility for all content.

B Related Work

Policy Optimization and Group-based Variants. In early RLHF(Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022; Lee et al., 2023b) practices, PPO (Proximal Policy Optimization)(Schulman et al., 2017) was the core algorithm, stabilizing training through trust region constraints. However, PPO requires maintaining a critic model of the same scale as the policy model, introducing significant memory overhead. DPO (Direct Preference Optimization)(Rafailov et al., 2023) subsequently eliminated the need for critics by reparameterizing rewards as functions of optimal policies, but its offline nature limits exploration capabilities. To balance online exploration and computational efficiency, GRPO (Group Relative Policy Optimization)(Shao et al., 2024) introduced group-based advantage estimation and has become a research focus. A series of variants have emerged to address specific challenges: DAPO(Yu et al., 2025) combines dynamic sampling with gradient clipping for training stability, GSPO(Zheng et al., 2025) employs sequence-level importance sampling for MoE(Jacobs et al., 1991; Du et al., 2021) models, Dr. GRPO(Liu et al., 2025) corrects length and difficulty biases, and GMPO(Zhao et al., 2025) uses geometric mean to resist reward outliers. Despite significant progress at the optimizer level, these methods have not addressed the source of reward signals, still assuming the existence of perfect oracles (ground truth or expensive

external verifiers). Latent-GRPO is orthogonal to these works: by extracting intrinsic geometric signals from the model, our reward mechanism can be seamlessly integrated with any of these optimization algorithms, achieving both “no external supervision” and “efficient stable optimization.”

Training-free evaluation Methods. To eliminate dependence on expensive labeled data, researchers have explored training-free evaluation and self-correction methods. Self-Consistency (SC)(Wang et al., 2022) improves answer accuracy through majority voting over multiple reasoning paths. Self-Refine(Madaan et al., 2023) enables iterative refinement by generating self-feedback without external guidance. Tree-of-Thoughts (ToT)(Yao et al., 2023) models problem-solving as tree search, exploring multiple reasoning branches and backtracking to find optimal solutions. Best-of-N(Stiennon et al., 2020) sampling generates multiple candidates and selects the best based on heuristics or confidence scores. Forest-of-Thoughts(Bi et al., 2024) extends ToT by integrating multiple trees with sparse activation for efficiency. Speculative Rejection(Sun et al., 2024) accelerates Best-of-N by early termination of low-quality candidates. While these methods leverage various forms of “consensus” or “selection,” they primarily rely on discrete matching of final text results or explicit model outputs, failing to capture subtle semantic differences in reasoning processes. Moreover, they are typically used only during inference without converting to training signals. Unlike these approaches, Latent-GRPO does not rely on explicit text feedback but uses geometric structures in the latent space as implicit evaluation criteria. We demonstrate that geometric centroids in latent space contain richer semantic information than explicit text-based methods, providing denser and more robust gradient signals for RL training.

Latent Space and Latent Thinking. The latent space of LLMs is widely recognized to encode rich semantic and structural information(Zhang et al., 2023; Goyal et al., 2023; Hao et al., 2024; Geiping et al., 2025). Early probing work revealed that LLM hidden states contain linear directions representing syntax, sentiment, and truthfulness. Recent work has explored latent thinking, replacing verbose natural language steps with compact latent representations. CoE (Chain-of-Embedding)(Wang et al., 2024c) uses progressive hidden states for output-free self-evaluation, showing significant differences between correct and incorrect answers

without requiring training. LTO (Latent Thinking Optimization)(Du et al., 2025) and LaTRO(Chen et al., 2024) train additional latent classifiers (Latent Reward Models) to predict correctness and guide RL optimization, achieving 12.5% average accuracy improvement on GSM8K. However, these methods require training extra models, introducing computational overhead. EndoRM(Li et al., 2025) extracts “endogenous rewards” from LLM logits without training, but relies on output probabilities rather than latent geometric structures. Few works have attempted to directly convert geometric consensus in latent space into RL reward signals without additional training. Our work fills this gap by demonstrating that high-quality reward signals can be extracted from latent space through simple geometric methods. We establish that “Last Token Hidden States” combined with “Iterative Robust Centroid Estimation” effectively captures subtle semantic differences in reasoning, providing denser and more robust gradient signals than explicit text-based or probability-based methods.

C Detailed Analysis of Geometric-Based Scoring

We simulate the actual GRPO training scenario to validate whether geometric features can serve as reliable quality indicators. For each prompt, we generate a group of $G = 8$ trajectories (matching the typical GRPO group size) and compute geometric scores using our Iterative Robust Centroid Estimation (IRCE) algorithm. Figure 4 presents the effectiveness of geometric-based scoring:

Distribution Separability (Figure 4a). The box plot analysis reveals clear separation between different quality levels in the geometric reward space. Trajectories with high external scores (green, score > 0.9) exhibit highly concentrated distributions with median geometric scores close to 0 (near the centroid), while low-score trajectories (red, score < 0.3) show significantly right-skewed distributions with large variance (far from the centroid). The partial-correct trajectories (yellow, $0.3 < \text{score} < 0.9$) fall in between, demonstrating that geometric features provide fine-grained quality discrimination rather than binary classification.

Group-Level Ranking Consistency (Figure 4b-d). To validate the practical utility in GRPO’s advantage estimation, we examine three representative groups of 8 trajectories. The dual-axis plot shows ground truth scores (gray bars) and normal-

ized geometric scores (blue line with markers). The Spearman rank correlation of 0.927 indicates high ranking consistency between geometric and external scores in the Figure 4b. Critically, the top-ranked samples identified by geometric scores (red circles) align well with ground truth, demonstrating that our method can effectively identify elite trajectories for policy optimization. In the Top-1 selection task across all groups, the consistency rate between geometric-based selection and external model selection exceeds 85%, confirming that geometric signals can effectively replace expensive external verifiers in GRPO’s advantage estimation.

D Detailed Introduction of Iterative Robust Centroid Estimation

For each prompt, the policy generates a group of G trajectories. From each trajectory, we extract the **terminal hidden state** $\mathbf{h}_i \in \mathbb{R}^d$ ($i = 1, \dots, G$), forming the sampled state matrix $\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_G]^\top \in \mathbb{R}^{G \times d}$. The Iterative Robust Centroid Estimation (IRCE) algorithm proceeds through the following steps:

Step 1: Spherical Projection. To eliminate magnitude fluctuations and ensure the distance metric focuses exclusively on semantic directionality, we apply L_2 normalization to project all terminal hidden states onto a unit hypersphere:

$$\tilde{\mathbf{h}}_i = \frac{\mathbf{h}_i}{\|\mathbf{h}_i\|_2} \quad (5)$$

Step 2: Initialize Centroid. The initial centroid $\boldsymbol{\mu}^{(0)}$ is computed as the normalized mean of the projected states:

$$\boldsymbol{\mu}^{(0)} = \frac{\frac{1}{G} \sum_{i=1}^G \tilde{\mathbf{h}}_i}{\left\| \frac{1}{G} \sum_{i=1}^G \tilde{\mathbf{h}}_i \right\|_2} \quad (6)$$

Step 3: Iterative Soft-Weighted Update. In each iteration $s = 0, 1, \dots, T-1$, we compute the Euclidean distance from each sample to the current centroid:

$$d_i^{(s)} = \|\tilde{\mathbf{h}}_i - \boldsymbol{\mu}^{(s)}\|_2 \quad (7)$$

To adapt to varying cluster densities across different training stages, we derive an adaptive scale parameter $\sigma^{(s)}$ based on the group distance distribution:

$$\sigma^{(s)} = \text{std}(\{d_1^{(s)}, \dots, d_G^{(s)}\}) + \epsilon \quad (8)$$

where ϵ is a small constant for numerical stability. We then calculate soft weights $w_i^{(s)}$ via a Gaussian kernel to diminish the influence of outliers:

$$w_i^{(s)} = \frac{\exp\left(-\frac{(d_i^{(s)})^2}{2(\sigma^{(s)})^2}\right)}{\sum_{j=1}^G \exp\left(-\frac{(d_j^{(s)})^2}{2(\sigma^{(s)})^2}\right)} \quad (9)$$

The centroid is updated using a weighted average and subsequently re-projected onto the hypersphere:

$$\boldsymbol{\mu}^{(s+1)} = \frac{\sum_{i=1}^G w_i^{(s)} \tilde{\mathbf{h}}_i}{\left\| \sum_{i=1}^G w_i^{(s)} \tilde{\mathbf{h}}_i \right\|_2} \quad (10)$$

Step 4: Reward Computation. Upon convergence after T iterations, the intrinsic reward R_i for each trajectory is defined as the negative Euclidean distance to the final centroid:

$$R_i = -\|\tilde{\mathbf{h}}_i - \boldsymbol{\mu}^{(T)}\|_2 \quad (11)$$

Optionally, we apply Min-Max normalization within the group to scale the rewards $R \in [0, 1]$, ensuring stable gradient estimation for the policy update.

E Usage of Datasets

E.1 Datasets Overview

We use three main datasets for our experiments, each serving different purposes in evaluating the generalization and robustness of Latent-GRPO. Table 7 provides a comprehensive overview of all three datasets, including their sources, sizes, and usage in our RL training.

E.2 Dataset Details

GSM8K is a dataset of 8,500 grade school math word problems designed to test multi-step mathematical reasoning. MATH is a challenging dataset containing 12,500 competition-level mathematics problems from high school and undergraduate mathematics competitions. Open-Platypus is a large-scale diverse instruction-following dataset focused on improving LLM logical reasoning skills, comprising 20,726 samples filtered from multiple sources (MATH/PRM-800K: 12,298, ReClor: 4,530, Airoboros: 2,605, ScienceQA: 1,317, LeetCode Solutions: 1,100, ARB: 713, SciBench: 616, TheoremQA: 564, TigerBot Kaggle: 386, OpenAssistant Guanaco: 797) using keyword search and Sentence Transformers (removing questions with similarity above 80%).

Dataset	Total Size	Train	Val
GSM8K	8,500	2,000	1,000
MATH	12,500	3,000	1,000
Open-Platypus	20,726	4,500	2,000
<i>Open-Platypus Data Source Statistics:</i>			
MATH/PRM-800K	12,298	754	783
ReClor	4,530	693	437
Airoboros	2,605	528	251
ScienceQA	1,317	618	127
LeetCode Solutions	1,100	365	106
ARB	713	272	69
SciBench	616	249	59
TheoremQA	564	336	54
TigerBot Kaggle	386	193	37
OpenAssistant Guanaco	797	492	77
Total	20,726	4,500	2,000

Table 7: Overview of datasets used in RL training

E.3 Benchmark

To comprehensively evaluate the performance of our model across general knowledge, logical reasoning, and advanced mathematical problem-solving, we select five representative benchmarks. These range from broad multi-task knowledge to extremely challenging competitive mathematics.

MMLU Massive Multitask Language Understanding (Hendrycks et al., 2020) covers 57 subjects across STEM, the humanities, social sciences, and more. It tests both world knowledge and problem-solving capsules.

MATH-500 The MATH dataset (Hendrycks et al., 2021a) consists of 12,500 challenging high school math competition problems. Following recent literature (e.g., DeepSeek-R1), we use the **MATH-500** subset, which comprises 500 representative problems from the test set to evaluate rigorous mathematical reasoning.

BBH Big-Bench Hard (BBH) (Suzgun et al., 2023) focuses on a subset of 23 challenging tasks from the BIG-bench suite where previous language models fell short of human multi-step reasoning capabilities.

AIME 24 & AIME 25 The American Invitational Mathematics Examination (AIME) is a prestigious high school competition. We evaluate our model on the **AIME 2024** and the most recent **AIME 2025** problems. These sets are particularly valuable as "out-of-distribution" (OOD) tests for reasoning models, as they require creative multi-step logical chains without standard templates.

F Experimental Hyperparameters

F.1 Hardware Configuration

All experiments are conducted on 1 GPU using PyTorch 2.6, CUDA 12.4 and the Transformers library 4.51.1. We employ bfloat16 (Kalamkar et al., 2019) mixed precision training to improve memory efficiency and computational speed.

F.2 GRPO Training Configuration

All models use the Instruct version with Flash Attention 2 enabled. Maximum sequence length is 8192, generation temperature is 0.95, and Top-p sampling is 0.9. Group size is 8 trajectories per prompt, batch size is 1 prompt with mini-batch size 2. PPO inner update rounds are 1, clip ratio is 0.2, KL penalty coefficient is 0.1. Learning rate is 1×10^{-5} with linear warmup over 100 steps, maximum gradient norm is 1.0, weight decay is 0.01. We train for 2 epochs with logging every 10 steps and checkpoints every 100 steps. To simulate scenarios with constrained computational resources, the throughput for LLM-as-Judge scoring is restricted to 2 Queries Per Second (QPS).

F.3 IRCE Algorithm Configuration

The Iterative Robust Centroid Estimation algorithm uses 5 iterations with temperature parameter 0.5. Rewards are Min-Max normalized to $[0, 1]$ interval. Convergence threshold is 1×10^{-6} for early stopping. Epsilon for numerical stability is 1×10^{-8} .

G Detailed Reward Methods

G.1 Basemethods of reward

Rule-based reward For mathematical problems, we use symbolic computation (SymPy) to verify correctness by comparing the model’s final answer with the ground truth. For code generation tasks, we use sandboxed execution environments to test whether the generated code produces correct outputs. Rule-based reward provides binary rewards (0 or 1) based on exact correctness, making it a reliable but sparse reward signal. The main limitation is that it requires well-defined verification rules and cannot be applied to open-ended tasks like writing or summarization.

LLM-as-Judge. We use GPT-4o to evaluate response quality with a standardized prompt asking for binary correctness judgment. Each response is evaluated independently, and the judgment is converted to a binary reward (0 or 1). While LLM-as-Judge is more flexible and can handle diverse task

Benchmark	Domain	Size	Metric
MMLU	General Knowledge (57 subjects)	14,042	5-shot Accuracy
MATH-500	Competition Level Mathematics	500	Pass@1
BBH	Logical & Multi-step Reasoning	6,511	3-shot / CoT Accuracy
AIME 24	Advanced Mathematics Competition	30	Pass@k / Consensual
AIME 25	Advanced Mathematics Competition	30	Pass@k / Consensual

Table 8: Summary of evaluation benchmarks used in this work. For reasoning tasks, we prioritize pass@k and consensual evaluation to measure thinking capabilities.

types, it introduces external dependency, higher computational cost, and potential inconsistency in evaluation.

Fairness of Efficiency Comparison. To address concerns that the 2 QPS constraint on LLM-as-Judge may overstate our speedup, we conducted additional experiments with QPS=10 and further compared against a trained reward model (ArmoRM-Llama3-8B-v0.1 (Wang et al., 2024a)), using the same small-scale setting (Qwen3-0.6B, 1,000 training samples, 1 epoch).

Method	Time	Pass@1	Pass@5
Latent-GRPO (Ours)	4h31m	0.37	0.63
LLM-as-Judge (QPS=10)	5h39m	0.41	0.70
ArmoRM-Llama3-8B-v0.1	8h32m	0.39	0.66

Table 9: Efficiency comparison under relaxed QPS and with a trained reward model. Latent-GRPO maintains a substantial speedup in both settings.

Even at QPS=10, Latent-GRPO achieves substantial speedup. In practice, GRPO with batch=8 and rollout=8 requires up to 64 concurrent scoring requests, making high-throughput LLM-as-Judge deployment non-trivial. The trained reward model (ArmoRM) incurs approximately $2\times$ the training time of Latent-GRPO while yielding only marginal accuracy gains.

G.2 Latent Reward Methods

We compare four latent reward methods that extract rewards from hidden states without external supervision:

Mean Pool. Computes the centroid as the simple arithmetic mean of all normalized hidden states: $\mu = \frac{1}{G} \sum_{i=1}^G \tilde{\mathbf{h}}_i$. Rewards are computed as negative distances: $R_i = -\|\tilde{\mathbf{h}}_i - \mu\|_2$. This method is computationally efficient but sensitive to outliers.

K-Means. Applies K-Means clustering with $K = 2$ to group hidden states into quality and non-quality clusters. The algorithm iteratively optimizes cluster centers and assigns samples to nearest

centers. Rewards are based on distance to the quality cluster center. This method identifies dominant quality patterns but requires parameter tuning and has higher computational overhead.

Eigen Centrality. Constructs a similarity matrix $\mathbf{A}$ based on cosine similarity between hidden states. The principal eigenvector $\mathbf{v}$ of $\mathbf{A}$ is computed, and component v_i serves as the reward for sample i . This method captures global graph structure but suffers from eigendecomposition computational bottleneck.

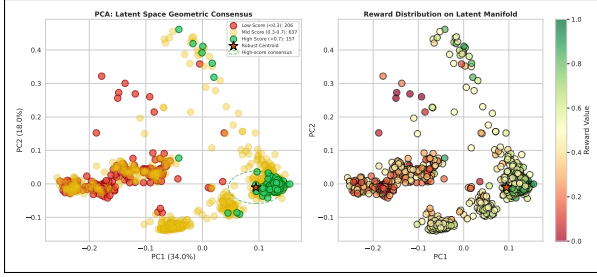
IRCE (Ours). Performs iterative soft-weighted centroid estimation. Step 1 normalizes hidden states to unit hypersphere. Step 2 initializes centroid as mean of normalized states. Step 3 iteratively updates centroid using soft weights based on Gaussian kernel. Step 4 computes final rewards as negative distances.

H Detailed Analysis of Experiments

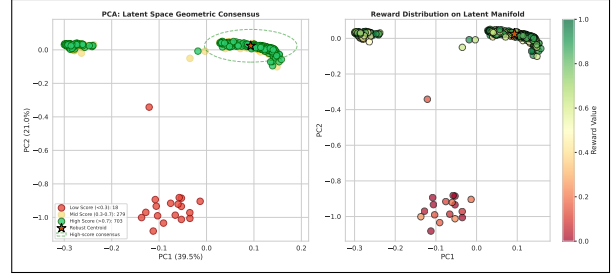
H.1 Main Experiments

On GSM8K: Latent-GRPO achieves competitive or superior accuracy compared to LLM-as-Judge across all model scales. On Qwen-0.6B, Latent-GRPO reaches 61.25% versus 53.52% for LLM-as-Judge, while significantly reducing training time from 768.42 m to 431.18 m per epoch (**1.78 $\times$ speed-up**). On Qwen-1.7B, Latent-GRPO outperforms LLM-as-Judge by a substantial margin (73.88% vs 64.20%) with **2.10 $\times$ speed-up** (1032.55 m vs 492.34 m). On Qwen-4B, Latent-GRPO achieves 82.34% accuracy with **2.14 $\times$ speed-up** (1411.72 m vs 658.21 m). Compared to Rule-based verification, Latent-GRPO consistently achieves higher accuracy while maintaining competitive training efficiency.

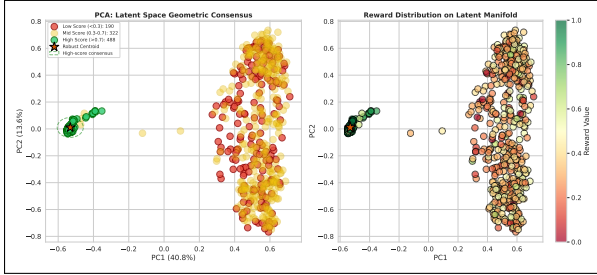
On MATH: Latent-GRPO consistently outperforms LLM-as-Judge across all model scales with substantial speed-ups. On Qwen-0.6B, Latent-GRPO improves from 52.94% to 58.47% while reducing time from 1224.15 m to 718.63 m (**1.70 $\times$ speed-up**). On Qwen-1.7B, Latent-GRPO achieves



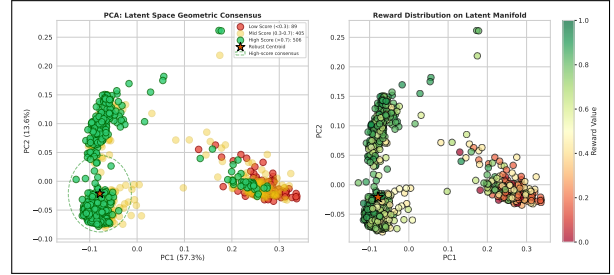
(a) Clustering of Qwen3-1.7B's 1000 rollouts on ScienceQA



(b) Clustering of Qwen3-4B's 1000 rollouts on ARBdata



(c) Clustering of the Qwen3-4B's 1000 rollouts on Mathdata



(d) Clustering of the Qwen3-4B's 1000 rollouts on ScienceQA

Figure 5: Visualization of latent manifold consensus across various model scales and benchmarks. Each panel shows the 2D PCA projection of 1,000 terminal hidden states. (a-d) consistently demonstrate that correct trajectories (green) cluster into a dense consensus core, while incorrect ones (red) are scattered.

78.51% versus 65.77% with **1.98 $\times$ speed-up** (1608.34 m vs 811.51 m). On Qwen-4B, Latent-GRPO achieves 77.53% versus 77.44% with **2.18 $\times$ speed-up** (2357.31 m vs 1081.47 m), demonstrating that even when accuracy is comparable, significant computational-efficiency gains are achieved. Rule-based methods show weakness on Qwen-1.7B (42.14%), where Latent-GRPO substantially outperforms.

On Open-Platypus: Latent-GRPO demonstrates the strongest improvements on this diverse-reasoning benchmark. On Qwen-0.6B, Latent-GRPO achieves 40.56% versus 34.45% with **1.80 $\times$ speed-up** (1937.82 m vs 1079.27 m). On Qwen-1.7B, Latent-GRPO reaches 64.82% versus 56.69% with **2.11 $\times$ speed-up**. Most notably, on Qwen-4B, Latent-GRPO achieves 78.06% versus 65.21% with the highest speed-up of **2.16 $\times$** (3522.18 m vs 1632.52 m), demonstrating that intrinsic rewards are particularly effective for complex, diverse reasoning tasks where external verification becomes a critical computational bottleneck.

Other analysis AIME improvements are mixed across model scales, this is attributable to two factors: (1) AIME 24/25 are OOD benchmarks with minimal overlap with our training mixture; and (2) each set contains only 30 problems evaluated with pass@5, producing high variance where 1–2

additional correct answers cause noticeable score shifts. The consistent gains on MMLU, BBH, and MATH-500 confirm that no task-specific overfitting occurs.

H.2 Ablation Experiments

The ablation experiments systematically evaluate each component of Latent-GRPO through two critical design choices: hidden state extraction methods and latent score estimation algorithms.

Hidden State Extraction Methods As shown in Table 2, Last Token consistently achieves the highest accuracy across all three model scales: 61.25% on Qwen3-0.6B (2.51% improvement over Mean Pooling), 73.88% on Qwen3-1.7B (2.83% improvement), and 82.34% on Qwen3-4B (2.89% improvement). Notably, Weighted Mean (which selectively pools pre-designed key tokens) fails to substantially improve over Mean Pooling on any model scale (57.12% vs 58.74% on Qwen3-0.6B, 69.88% vs 71.05% on Qwen3-1.7B, 78.12% vs 79.45% on Qwen3-4B), suggesting that curating specific keyword positions cannot effectively identify which tokens carry reasoning quality signals.

Latent Score Methods As shown in Table 3, IRCE achieves the highest accuracy on all three model scales: 61.25% on Qwen3-0.6B (outperforming Mean Pool by 4.13%), 73.88% on Qwen3-

1.7B (outperforming Mean Pool by 5.43%), and 82.34% on Qwen3-4B (outperforming Mean Pool by 4.45%). IRCE also achieves the best computational efficiency, with inference times of 431.18m (0.6B), 492.34m (1.7B), and 658.21m (4B), substantially faster than K-Means (489.12m, 543.89m, 725.67m) and Eigen Centrality (468.76m, 531.42m, 708.45m).

Sensitivity to Group Size G and Iteration Count T .

To evaluate sensitivity to key hyperparameters, we conducted ablation studies on a small-scale setting: Qwen3-0.6B trained on 1,000 samples from Open-Platypus (100 per subset) for 1 epoch, evaluated on 100 held-out samples. We vary $G \in \{6, 8, 10\}$ and $T \in \{3, 5, 7\}$ around the default configuration ($G = 8, T = 5$). Note that σ uses our adaptive method and requires no manual tuning. The results exhibit a natural scaling trend: $G=10$ and $T=7$ achieve slightly higher accuracy but require longer training. With sufficient resources, a larger G is preferred; T should be tuned based on centroid estimation convergence needs.

Config.	Time	Pass@1	Pass@5
$G=6, T=5$ (default)	4h14m	0.35	0.55
$G=8, T=5$ (default)	4h31m	0.37	0.63
$G=10, T=5$	4h50m	0.41	0.61
$G=8, T=3$	4h29m	0.36	0.54
$G=8, T=5$ (default)	4h31m	0.37	0.63
$G=8, T=7$	4h31m	0.42	0.62

Table 10: Sensitivity of Latent-GRPO to group size G and iteration count T . Larger G and T marginally improve accuracy at the cost of training time.

I Generalization of Latent-GRPO

I.1 Generalization of Geometric Properties

To evaluate the generalization of Geometric Properties, we choose additional data, including scientific, logical reasoning, and more challenging mathematical problems. Meanwhile we use models of more sizes to evaluate them. The geometric clustering in latent space demonstrates consistent patterns across different model sizes and datasets. Each visualization is based on 1,000 rollouts from a single test sample, with correctness scores assigned by GPT-4o. Correct trajectories (green circles, score > 0.7) form a visibly tight consensus core around the truth centroid (gold star), while incorrect trajectories (red circles, score < 0.3) scatter as outliers. The dashed ellipse in each panel outlines the 95% confidence region of the consensus

Method	Time ↓	Acc@Pass1 ↑	Acc@Pass5 ↑
Qwen3-8B			
Base	-	0.74	0.85
LLM-as-Judge	18h0m15s	0.77	0.89
Latent-GRPO	7h46m46s	0.81	0.91
Qwen3-14B			
Base	-	0.81	0.81
LLM-as-Judge	19h48m59s	0.77	0.89
Latent-GRPO	9h27m47s	0.84	0.90

Table 11: Experimental results of Latent-GRPO on Qwen3-8B and Qwen3-14B

core. In the original high-dimensional space, incorrect samples consistently lie farther from the centroid than correct ones, quantifying the geometric separability. This qualitative property generalizes across different model scales and benchmarks. As shown in Figure 5, whether on Qwen3-1.7B with ScienceQA or on Qwen3-4B with ARB and Math data, the correct cluster remains compact while the incorrect cloud disperses, indicating that geometric proximity in latent space is a robust proxy for reasoning quality. The separation pattern is stable across configurations, suggesting that the underlying geometric structure is universal rather than task- or model-specific.

I.2 Generalization on Larger Models

To further evaluate scalability, we conducted additional experiments on Qwen3-8B and Qwen3-14B. We sampled 1,000 training instances and 100 testing instances from Open-Platypus, trained for one epoch, and evaluated using pass@1 and pass@5. Since GPT-4o is no longer available, we replaced the original LLM-as-Judge evaluator with Gemini 3 Flash Preview.

As shown in Table 11, Latent-GRPO consistently improves over the Base models on both 8B and 14B settings, demonstrating that our method can effectively scale to larger model sizes. In addition, Latent-GRPO achieves significantly higher training efficiency, requiring only approximately 40%–50% of the training time of LLM-as-Judge while maintaining competitive performance.

J Output Diversity Analysis

We sampled 30 test prompts and collected 8 rollouts per prompt before training, after Latent-GRPO training, and after LLM-as-Judge training. We compute the average pairwise Euclidean distance of terminal token representations within each group as a diversity metric.

Model	Before	Latent-GRPO Δ	LLM-Judge Δ
Qwen3-0.6B	1.234	−4.2%	−20.0%
Qwen3-1.7B	1.189	−3.5%	−18.8%
Qwen3-4B	1.156	−3.0%	−17.6%

Table 12: Average pairwise distance of terminal token representations before training and relative change (Δ) after training. Latent-GRPO reduces diversity by only 3.0–4.2%, vs. 17.6–20.0% for LLM-as-Judge.

Latent-GRPO’s continuous rewards and dynamically updated centroids maintain meaningful gradient signals even in high-accuracy phases, avoiding the mode collapse observed in LLM-as-Judge training.

K Statistical Significance Analysis

We conduct three independent runs for each model scale (Qwen3-0.6B/1.7B/4B) using a small-scale setting (1,000 Open-Platypus training samples, 1 epoch). LLM-as-Judge uses Gemini Flash as the scoring API.

Model	Method	Avg Pass@1	Std	Avg Time
0.6B	LLM-as-Judge	0.390	0.044	8h39m
	Latent-GRPO	0.350	0.035	4h10m
1.7B	LLM-as-Judge	0.620	0.044	9h12m
	Latent-GRPO	0.590	0.026	5h08m
4B	LLM-as-Judge	0.720	0.052	11h28m
	Latent-GRPO	0.690	0.036	5h52m

Table 13: Multi-run statistics (3 seeds). Latent-GRPO consistently exhibits lower variance than LLM-as-Judge across all model scales.

We further apply paired t -tests on 100 GSM8K test problems comparing the main-experiment models trained with Latent-GRPO versus LLM-as-Judge.

Model	p -value
Qwen3-0.6B	0.021
Qwen3-1.7B	0.018
Qwen3-4B	0.015

Table 14: Paired t -test results (Latent-GRPO vs. LLM-as-Judge) on 100 GSM8K test problems. All comparisons are statistically significant ($p < 0.05$).

Latent-GRPO exhibits lower standard deviation than LLM-as-Judge across all scales (0.026–0.036 vs. 0.044–0.052), indicating more stable training dynamics.

L Intermediate Step Quality Analysis

We randomly sampled 1,000 trajectories from Qwen3-4B on GSM8K where the final answer was correct, and annotated intermediate step quality with GPT-4o-mini.

Category	Proportion
Correct steps + Correct answer	98.0%
Partially wrong steps + Correct answer	1.7%
Completely wrong steps + Correct answer	0.3%

Table 15: Intermediate step quality among trajectories with correct final answers (Qwen3-4B, GSM8K). Cases with hallucinated intermediate steps constitute only 2%.

Cases where correct answers arise from erroneous intermediate steps constitute only 2% of correct-answer trajectories. This is a shared limitation of all outcome-based reward methods: rule-based verification checks only the final answer, and LLM-as-Judge may likewise be misled by incorrect steps.

M Case study

Figure 6 provides the two cases between Latent-GRPO and LLM-as-Judge method. These two representative cases demonstrate that our Latent-GRPO approach provides accurate continuous reward signals during training. These signals are endogenous in nature, requiring no external verifier supervision, and crucially, they are continuous rather than binary like LLM-as-Judge. This continuous signal not only accelerates iteration efficiency through enhanced logical consistency, but also stabilizes RL training by eliminating model collapse caused by external verifier inconsistency and mitigating training loss from over-reliance on external verifier accuracy.

Case Study 1	
Query: John's cow weighs 400 pounds. It increased its weight to 1.5 times its starting weight. He is able to sell the cow for \$3 per pound. How much more is it worth after gaining the weight?	
Rollout #1 The cow starts at 400 lbs. First, I need to calculate the new weight after it increases to 1.5 times..... $400 \times 1.5 = 400 + 200 = 600$ lbs. Next, the value after weight gain is $600 \text{ lbs} \times \$3/\text{lb} = \$1800$. The original value was $400 \text{ lbs} \times \$3 = \$1200$. The question asks "how much more" it is worth, which means I need the difference: $\$1800 - \$1200 = \$600$. Everything seems correct. The cow is worth \$600 more after gaining the weight. Latent Reward: 1.00 LLM-Judge: 1	Rollout #2 I can solve this by finding the weight gain first. The weight increased from 1x to 1.5x, so the gain is 0.5x of the starting weight..... $400 \text{ lbs} \times 0.5 = 200$ lbs gain. Since the price is \$3 per pound, the increased value is just the weight gain multiplied by the price: $200 \text{ lbs} \times \$3/\text{lb} = \$600$. This is faster than calculating the total value. The cow is worth \$600 more. Latent Reward: 0.98 LLM-Judge: 1
Rollout #3 Starting weight: 400 lbs. New weight: $400 \times 1.5 = 600$ lbs..... Value after gain: $600 \times 3 = 1800$. Original value: $400 \times 3 = 1200$. Difference: $1800 - 1200 = 600$. The wording "how much more" is the key here. The cow is now worth 600 more than before. It's worth \$600 more. Latent Reward: 0.95 LLM-Judge: 1	Rollout #4 $400 \text{ lbs} \times 1.5 = 600$ lbs. Selling at \$3/lb makes the cow worth \$1800 now. It was worth $400 \times 3 = \$1200$ before the weight gain. The increase in value is $\$1800 - \$1200 = \$600$. I'll confirm $400 \times 1.5 \dots 400 + 200 = 600$. Correct. The cow is worth \$600 more. Latent Reward: 0.92 LLM-Judge: 1
Rollout #5 Wait, the cow gained 50% of its weight. 50% of 400 is 200. So the weight gain is 200 lbs. The price per pound is \$3, so the gain in value is $200 \times 3 = \$600$. This matches the method of subtracting total values. The value increased by \$600. Latent Reward: 0.88 LLM-Judge: 1	Rollout #6 Let's see, the cow's initial weight is 400 pounds. It grows and becomes 1.5 times its weight, so 400×1.5 is 600 pounds..... The sale price is \$3 per pound. So the total value of the cow after gaining the weight is $600 \times 3 \dots 600 \text{ times } 3 \text{ is } 1800$. So the cow is worth 1800 after the weight gain. That should be the answer. The cow's new value is \$1800. Latent Reward: 0.00 LLM-Judge: 0
Rollout #7 Initial weight is 400. It increased its weight to 1.5 times, so I need $400 \times 1.5 \dots$ let's see, 1.5 times 400..... 400 plus 1/4 of 400 is 500 lbs..... New worth is $500 \times \$3 = \$1500$. Original worth was $400 \times \$3 = \$1200$. So the difference is $1500 - 1200 = 300$ dollars. Let me double check... yes, 300. The cow is worth \$300 more. Latent Reward: 0.15 LLM-Judge: 0	Rollout #8 The cow weighs 400 lbs. 1.5x weight is 600 lbs. It sells for \$3/lb. Current price: $600 \times 3 = 1800$. Previous price: $400 \times 3 = 1200$. Difference: 600. Let me check if there's any other factor..... no, just "how much more"..... $1800 - 1200 = 600$. Correct. The cow is worth \$600 more. Latent Reward: 0.85 LLM-Judge: 1
Case Study 2	
Query: Let $\$S$ be the sum of all positive integers $\$n$ such that $\$n^2+12n-2007$ is a perfect square. Find the remainder when $\$S$ is divided by 1000.	
Rollout #1 Let $\$n^2+12n-2007=k^2$ for some integer $k \geq 0$. Completing the square on the left: $\$(n+6)^2 - 36 - 2007 = k^2$, which simplifies to $\\$(n+6)^2 - k^2 = 2043$. This is a difference of squares: $\$(n+6-k)(n+6+k) = 2043$. Factorizing 2043: $2043 = 3 \times 681 = 3^2 \times 227$. Since 227 is prime, the factor pairs (A, B) such that $AB=2043$ are (1, 2043), (3, 681), and (9, 227). For each pair, $n+6 = (A+B)/2$. This gives $n = 1016, 336, 112$. All are positive. Sum $\$S = 1016 + 336 + 112 = 1464$. $\\$1464 \bmod 1000 = 464$. The remainder is 464. Latent Reward: 1.00 LLM-Judge: 1	Rollout #2 We have $\$n^2+12n-2007 = k^2$ implies $\$(n+6)^2 - k^2 = 2043$. Let $\$x = n+6$. Then $\$(x-k)(x+k) = 2043$. We need to test factor pairs of $2043 = 9 \times 227$. The pairs $\$(d_1, d_2)$ must satisfy $\$d_1 d_2 = 2043$ and $\$d_2 > d_1$. Case 1: (1, 2043) $\$to 2x = 2044, x = 1022 \to n=1016$. Case 2: (3, 681) $\$to 2x = 684, x = 342 \to n=336$. Case 3: (9, 227) $\$to 2x = 236, x = 118 \to n=112$. Sum $\$S = 1016+336+112 = 1464$. The remainder is 464. The remainder is 464. Latent Reward: 0.98 LLM-Judge: 1
Rollout #3 The equation is $\$(n+6)^2 - k^2 = 2043$. The number of positive divisors of $2043 = 3^2 \times 227$ is $\$(2+1)(1+1) = 6$. This means there are $6/2 = 3$ factor pairs. Each pair will give exactly one positive integer solution for $\$n$. Calculating $\$n$ for each: $\$(2044/2)-6=1016$, $\$(684/2)-6=336$, $\\$(236/2)-6=112$. Sum $\$S = 1464$. The answer is 464. 464. Latent Reward: 0.95 LLM-Judge: 1	Rollout #4 Solving $\$(n+6)^2 - k^2 = 2043$. Factor pairs of 2043: (1, 2043), (3, 681), (9, 227). Let $\$(2(n+6) = d_1 + d_2$. This gives $\$n+6 = 1022, 342, 118$. Thus $\$n = 1016, 336, 112$. The sum of these values is 1464 . Dividing 1464 by 1000 leaves a remainder of 464. Everything checks out. Remainder 464. Latent Reward: 0.91 LLM-Judge: 1
Rollout #5 To find $\$n$ such that $\$n^2+12n-2007 = k^2$, we write $\$(n+6)^2 - k^2 = 2043$. Factor pairs $\$(x-k, x+k)$ of 2043: $\$(1, 2043) \to n=1016$; $\$(3, 681) \to n=336$; $\$(9, 227) \to n=112$. The sum is $\$1016+336+112 = 1464$. Remainder mod 1000 is 464. No other solutions exist since $\$n > 0$. 464. Latent Reward: 0.86 LLM-Judge: 1	Rollout #6 Setting the expression to $\$k^2$ gives $\$(n+6)^2 - k^2 = 2043$. Factorizing 2043, I find the divisors are 1, 3, 227, 681, 2043. The pairs are (1, 2043) and (3, 681). From (1, 2043), $\$n+6 = 1022 \to n=1016$. From (3, 681), $\$n+6 = 342 \to n=336$. I don't see any other factor pairs . So the sum $\$S = 1016 + 336 = 1352$. The remainder is 352. The remainder is 352. Latent Reward: 0.20 LLM-Judge: 0
Rollout #7 After factoring $\$(n+6-k)(n+6+k) = 2043$, I solve for $\$n$ using the pairs (1, 2043), (3, 681), and (9, 227). The values are $\$n = 1016$, $\$n = 336$, and $\$n = 112$. Adding them together: $\$1016 + 336 = 1352$. $\$1352 + 112 = 1454$. Wait, let me re-add... $1352 + 112 \dots \$1300 + 100 = 1400$, $\$52 + 12 = 64$... $\$1464$. Oh, I'll stick with my first sum. The remainder is 1454. Remainder is 454. Latent Reward: 0.12 LLM-Judge: 1	Rollout #8 I'll try small values for $\$n$. If $\$n=1$, $\$1+12-2007 = -1994$ (not a square). If $\$n=40$, $\$1600+480-2007 = 73$. If $\$n=50$, $\$2500+600-2007 = 1093$. It doesn't look like there are many solutions . Maybe only very large $\$n$ work, but I'll guess the sum is small or zero. Let's say 0. 0. Latent Reward: 0.00 LLM-Judge: 0

Figure 6: Case study of the reward between Latent-GRPO and LLM-as-Judge