SDAE: Semantic-Diversity-Aware Exploration for Efficient Reinforcement Learning in Large Language Models

Jianwen Sun[1,2,3] sunjw@ccnu.edu.cn
Wangzi Shi[2,3] wangzi@mails.ccnu.edu.cn
Sannyuya Liu[1,2,3] liusy027@ccnu.edu.cn
Luona Wei[4] wlnelysion@scuec.edu.cn
Zhiming Wang[5] wang-zm18@mails.tsinghua.edu.cn
Fengpeng Yue[5] 11930381@mail.sustech.edu.cn
Qian Wan[1,2,3*] wanq8228@ccnu.edu.cn

[1] Laboratory for Artificial Intelligence and New Forms of Education, Central China Normal University
[2] Faculty of Artificial Intelligence in Education, Central China Normal University
[3] National Engineering Research Center of Educational Big Data, Central China Normal University
[4] College of Electronics and Information Engineering, South-Central Minzu University
[5] Independent Researcher

Abstract

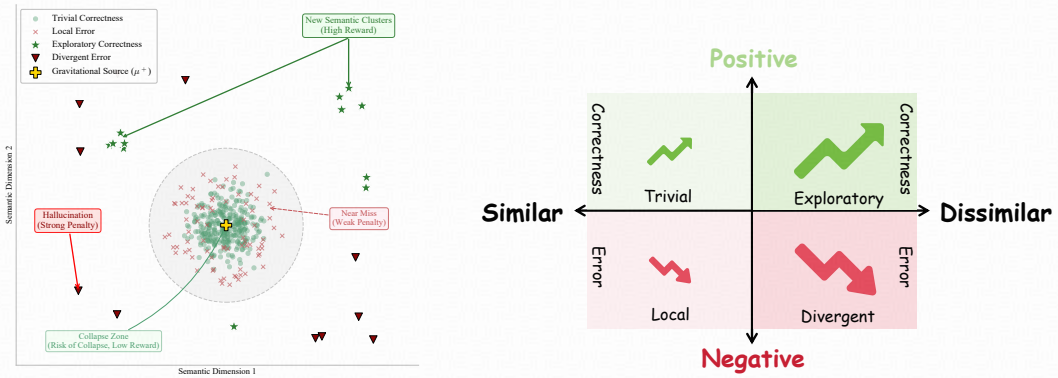
Reinforcement learning with verifiable rewards (RLVR) significantly enhances the reasoning capabilities of large language models. However, as training progresses, models often converge to narrow solution paths, leading to diversity collapse. Existing methods alleviate this phenomenon through token-level entropy regularization or negative sample reinforcement, yet these works have treated responses as atomic units distinguished only by binary correctness labels, overlooking substantial semantic heterogeneity among responses under the same label—among incorrect responses, the degree and type vary significantly; among correct responses, both conventional and novel strategies exist. Uniformly reinforcing all correct responses leads to policy convergence, while uniformly penalizing all incorrect ones may suppress promising directions. Building on this insight, we propose **S**emantic **D**iversity-**A**ware **E**xploration (**SDAE**), which leverages geometric structures in semantic space to guide exploration. SDAE computes the semantic centroid of positive samples and modulates advantage functions based on each response's distance to this centroid, encouraging novel correctness, penalizing divergent errors, while preserving improvement potential in locally flawed responses. Experiments demonstrate that SDAE consistently outperforms strong baselines across competition-level reasoning benchmarks, achieving a 13.3% improvement in Pass@256 on AIME25. Our code is available at `https://github.com/TTTigger/SDAE`.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) has emerged as the dominant paradigm for enhancing reasoning capabilities in large language models (LLMs) [1, 2, 3], with its core mechanism

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) This figure illustrates how SDAE categorizes and modulates responses within the latent manifold of the reasoning space.

(b) Classification of SDAE feedback. Arrow colors denote correctness (green) or error (red), while thickness represents reward/penalty magnitude.

Figure 1: The Semantic Geometry of Exploration.

reinforcing correct responses and penalizing incorrect ones through policy gradient algorithms in a simple yet effective manner [4, 2]. However, this purely accuracy-driven optimization objective harbors a fundamental tension: as training progresses, models tend to converge toward narrow solution paths, gradually losing the impetus to explore alternative reasoning strategies. This phenomenon, commonly referred to as diversity collapse, manifests as a sustained decline in the entropy of output distributions and stagnation or even degradation in performance on complex reasoning tasks [5, 6, 7].

In traditional reinforcement learning, entropy serves as a fundamental metric for exploration, quantifying the uncertainty in a model's action distribution. To address diversity collapse, existing research has primarily proceeded along two lines. The first line follows maximum entropy reinforcement learning principles, maintaining exploration capability by encouraging policy entropy [8, 9]. Recent work has demonstrated positive correlations between high-entropy regions and exploratory reasoning actions [10, 11]. The fundamental limitation of such methods lies in their operation at the token level, measuring uncertainty in lexical choices rather than semantic content—models may exhibit high entropy through surface phrasing variations while repeatedly exploring the same underlying strategies. The second line employs negative sample reinforcement (NSR), penalizing only incorrect responses to maintain generation diversity [12, 13]. NSR effectively prevents model overconfidence on correct samples by suppressing erroneous generations and redistributing probability mass to reasonable alternatives. However, NSR treats all incorrect responses uniformly, regardless of whether they represent promising near-miss directions or fundamental errors, while also failing to actively guide models toward exploring semantically diverse correct solutions.

Through in-depth analysis of the training process, we observe a critical granularity deficiency in how existing methods handle response "quality." Taking mathematical reasoning as an example: among incorrect responses, there exist "near-miss" local failures and "completely off-track" fundamental errors; among correct responses, there exist "clichéd" conventional paths and "innovative" novel strategies. The binary labels universally adopted by existing methods obliterate these crucial nuances, failing to identify "near-miss" samples with improvement potential or to provide additional incentives for genuinely exploration-worthy novel solutions. This signal deficiency is precisely what causes models to fall into diversity collapse during late training stages, struggling to break through reasoning bottlenecks.

We discover that this hidden quality differential exhibits remarkably clear geometric manifestation in semantic embedding spaces. As illustrated in Figure 1a, when mapping model responses to latent space manifolds, responses of varying qualities exhibit distinctive topological signatures on the latent manifold: semantically similar conventional solutions spontaneously cluster around the centroid of correct solutions, while novel solutions and divergent errors distribute in peripheral regions far from the centroid. This discovery reveals a key insight: **the geometric distance from a response to the semantic centroid essentially encodes its marginal value for model exploration.** Based on this, we construct a four-category semantic geometry framework: Trivial Correctness (conventional solutions), Exploratory Correctness (novel paths), Local Error (near-miss), and Divergent Error (fundamental

failures). This perspective exposes the common blind spot of existing methods: neither token-level entropy nor binary labels exploit the geometric information embedded in semantic space.

Building on these insights, we propose Semantic-Diversity-Aware Exploration (SDAE), a framework that guides exploration through semantic space geometry rather than token-level entropy. Its core logic lies in using the semantic centroid of correct responses as a "gravitational source," measuring each response's distance to this centroid (termed Semantic Gravity), and differentially modulating advantage functions accordingly. Specifically, SDAE comprises three steps: first, leveraging the policy model's own hidden states to perform semantic aggregation on sampled responses, with the cluster center of correct responses serving as the gravitational source; second, computing each response's semantic gravity to the centroid and combining it with correctness labels to classify samples into four categories; finally, establishing tiered advantage modulation—amplifying rewards for novel correct paths, strengthening penalties for divergent errors, and attenuating penalties for local errors to preserve improvement potential. Experiments demonstrate that SDAE consistently outperforms strong baselines across multiple benchmarks, improving Pass@256 from 50.0% to 63.3% on AIME25 while effectively mitigating entropy collapse.

The main contributions of this work are as follows:

- We reveal the structural problem of binary reward signals losing semantic geometric information in RLVR, and show that correctness labels and semantic distance capture complementary, non-redundant information about a response's exploration value. Based on this, we propose a semantic geometry-based four-category framework, providing a new perspective for understanding diversity collapse.
- A simple yet effective solution, SDAE: a semantic geometry-guided advantage modulation method. By measuring each response's semantic proximity to the positive centroid, SDAE achieves differentiated advantage modulation across four categories, adaptively balancing exploration and exploitation without introducing additional hyperparameters.
- Extensive experiments validate SDAE's effectiveness: consistently outperforming strong baseline methods across multiple competition-level reasoning benchmarks, achieving up to 13.3% absolute improvement in Pass@256 on AIME25, demonstrating powerful reasoning scaling capabilities.

2 Related Work

2.1 Reinforcement Learning for LLM reasoning

Reinforcement learning (RL) has become a core mechanism for enhancing the reasoning and decision-making capabilities of LLMs, extending their abilities beyond supervised fine-tuning [14, 15, 16]. Recently, RLVR has emerged as a promising paradigm for enhancing LLM reasoning capabilities, particularly in math and coding [2, 1, 17]. Unlike traditional RLHF, which relies on reward models trained on human preferences, RLVR employs deterministic verification functions to evaluate response correctness. This simple yet effective mechanism offers multiple advantages: mitigating reward hacking by providing ground-truth feedback, eliminating the need for extensive human annotation and complex reward model training, and demonstrating significant sample efficiency [12]. OpenAI o1 [18] demonstrates that RL can effectively incentivize reasoning capabilities at scale, followed by models such as DeepSeek-R1 [1], QwQ [19], Kimi k1.5 [20], and Qwen3 [21].

2.2 Exploration Strategies in Reinforcement Learning

Exploration strategies in RLVR have evolved through multiple paradigms. Entropy-based methods prevent policy collapse by adding entropy rewards, but RLVR often leads to entropy collapse, where probability mass concentrates on limited tokens [8, 22, 23, 24]. Research shows that positive advantage tokens are the primary contributors, with most entropy consumption occurring in early training stages [7, 11, 25]. Negative sample reinforcement (NSR) provides an alternative, where training solely on negative samples can improve Pass@k by suppressing incorrect responses and redistributing probability mass to reasonable alternatives [12, 13]. Semantic-level diversity methods have recently emerged, including EVOL-RL's novelty-aware rewards based on semantic dissimilarity [26], DARLING's learned binary classifier for diversity measurement [23], G^{2}RL's first-order update geometry

revealing external semantic embeddings as unreliable exploration proxies [25], and SEED-GRPO's semantic entropy regularization [27]. Unlike these prior methods that rely on token-level entropy or polarity-based advantage shaping, we propose SDAE, which leverages geometric structure in continuous semantic space for finer-grained exploration guidance.

3 Method

3.1 Motivation: The Semantic Geometry of Exploration

The core insight of our method is that model responses to the same question occupy different positions in semantic embedding space, with their geometric relationships encoding rich information about reasoning diversity. Standard RLVR methods apply uniform rewards to all correct responses and uniform penalties to all incorrect responses, ignoring this semantic structure. However, we argue that this structure should be leveraged to guide exploration in RL.

We propose the centroid of correct responses $\boldsymbol{\mu}^+$ as the gravitational source. Each response's distance to this gravitational source—termed Semantic Gravity (SG)—captures its degree of deviation from the "correctness center." This naturally produces distinctive topological signatures in the latent manifold, as visualized in Figure 1a. We formalize this observation into a four-category system (Figure 1b) that reveals the semantic geometry of exploration:

- **Trivial Correctness:** Correct responses clustered near the centroid represent conventional solution paths the model has already mastered. Reinforcing these is necessary, but over-rewarding risks diversity collapse.
- **Exploratory Correctness:** Correct responses far from the centroid represent novel, semantically diverse solution paths. The model has discovered a novel yet effective solution path; these should receive amplified rewards to encourage continued semantic diversification.
- **Local Error:** Incorrect responses still semantically close to correct solutions represent "near-misses." Strongly penalizing these would suppress promising reasoning directions; instead, weak penalties allow the model to refine these approaches.
- **Divergent Error:** Incorrect responses semantically distant from correct solutions represent fundamental reasoning failures. These should receive strong penalties to prevent the model from wasting capacity on ineffective reasoning directions.

This semantic geometry perspective enables SDAE to dynamically calibrate advantage values based on both correctness and semantic distance, balancing exploration and exploitation in a semantically meaningful way.

3.2 SDAE:Semantic-Diversity-Aware Exploration

The SDAE framework achieves dynamic reshaping of reinforcement learning advantage functions by perceiving the geometric topology of samples in latent space. Its core logic lies in using semantic gravity as a modulation factor to explicitly balance the model's "exploitation" on known high-frequency paths with "exploration" toward unknown semantic boundaries. As shown in Figure 2, the algorithm execution flow divides into three standardized phases:

Phase 1: Rollout Sampling. For a given question q, we first sample N responses $\mathcal{Y} = \{y_1, \ldots, y_N\}$ from the current policy π_θ.

Phase 2: Latent Semantic Aggregation & Semantic Gravity Calculation. To ensure that semantic representations remain aligned with the policy model's internal reasoning dynamics, we employ Latent Semantic Aggregation (LSA), which directly leverages the last-layer hidden states of the policy model. This design is theoretically motivated by recent findings demonstrating that the final layer serves as the unique optimization bottleneck and that external encoders exhibit geometric misalignment with the policy's gradient space [25]. We provide a detailed discussion in Appendix A.

Let $\mathbf{h}_{i,t}$ denote the last-layer hidden state of the i-th rollout at time step t. Its global semantic vector v_i is defined as:

$$v_i = \text{MeanPool}(\{\mathbf{h}_{i,t}\}_{t=1}^{T_i}) \in \mathbb{R}^d. \tag{1}$$

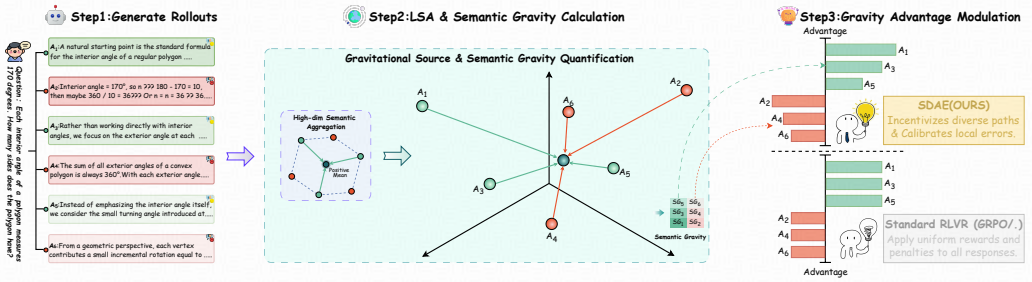


Figure 2: Illustration of the SDAE framework. The process encompasses LLM rollout generation, semantic gravity calculation, and advantage modulation. Unlike GRPO's uniform reward scheme, SDAE enables differentiated reinforcement guided by semantic geometry, encouraging diverse correct paths while suppressing divergent errors.

Simultaneously, based on verifiable rewards $r_i \in \{+1, 0\}$, we partition samples into correct and incorrect sets:

$$\mathcal{Y}^+ = \{y_i : r_i = +1\}, \mathcal{Y}^- = \{y_i : r_i = 0\}. \tag{2}$$

Let $\mathcal{P} = \{y_i : r_i = 1\}$ denote the set of correct responses. The gravitational source is:

$$\boldsymbol{\mu}^+ = \frac{1}{|\mathcal{P}|} \sum_{y_i \in \mathcal{P}} v_i. \tag{3}$$

Subsequently, we compute the semantic gravity for each sample y_i, i.e., its degree of deviation from the consensus path on the cosine manifold:

$$SG_i = 1 - \frac{v_i \cdot \mu^+}{\|v_i\| \|\mu^+\|}. \tag{4}$$

Smaller SG_i indicates the response is semantically closer to the centroid, while larger SG_i indicates semantic deviation.

Phase 3: Four-Category Advantage Modulation. Unlike the uniform rewards applied by standard GRPO [2], SDAE introduces the modulation coefficient SG to reweight the original advantage function A_i (see Appendix E for the complete GRPO formulation):

$$A_i^{SDAE} = A_i \cdot \hat{SG}_i. \tag{5}$$

The SDAE-modulated GRPO objective then becomes:

$$\mathcal{J}_{SDAE}(\theta) = \mathbb{E}_{[q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(O|q)]} \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}\hat{A}_{i,t}^{SDAE},\right.\right.$$
$$\left.\left.\text{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}^{SDAE}\right) - \beta\mathbb{D}_{KL}[\pi_\theta||\pi_{ref}]\right\}. \tag{6}$$

This simple linear reweighting mechanism mathematically and automatically implements the four-category exploration logic we propose. See Appendices B-D for further implementation details.

3.3 Theoretical Justification of SDAE

To delve deeper into SDAE's internal mechanism, we analyze how semantic gravity modulates policy gradient updates to elucidate its logic for balancing exploration and exploitation in semantic manifolds. For clarity, consider the loss function without clipping:

$$L_i(\theta) = \text{ratio}_i(\theta) \cdot \hat{A}_i = \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)} \cdot \hat{A}_i. \tag{7}$$

The gradient is computed as:

$$\nabla_\theta L_i(\theta) = \nabla_\theta \log \pi_\theta(o_i|q) \cdot \text{ratio}_i(\theta) \cdot \hat{A}_i. \tag{8}$$

Correspondingly, the policy update becomes (with global learning rate η):

$$\theta \leftarrow \theta + \eta \cdot \nabla_\theta \log \pi_\theta(o_i|q) \cdot \text{ratio}_i(\theta) \cdot \underbrace{A_i \cdot \hat{SG}_i}_{\hat{A}_i}. \tag{9}$$

As shown above, policy updates are controlled by four components: the global learning rate η, the log-probability gradient $\nabla_\theta \log \pi_\theta(o_i|q)$, the importance sampling ratio $\text{ratio}_i(\theta)$, and the semantically-aware advantage $\hat{A}_i = A_i \cdot \hat{SG}_i$. By integrating the semantic gravity factor, SDAE effectively modulates update magnitudes according to each response's geometric position in semantic space. This can be viewed as dynamically adjusting the effective learning rate at the response level: the original advantage A_i determines the direction of parameter updates (reinforcement or suppression), while $\hat{SG}_i$ scales the magnitude according to semantic distance from the positive centroid.

This design produces an important property: responses semantically close to known correct solutions receive attenuated gradients regardless of their correctness labels. For correct responses, this avoids redundant reinforcement of already well-captured patterns. For incorrect responses, this preserves exploration potential in promising semantic regions, as minor adjustments may lead to correct solutions. Conversely, responses far from the positive centroid receive amplified gradients, enabling the model to strongly reinforce novel correct reasoning while decisively suppressing fundamentally flawed paths. This adaptive modulation creates an implicit curriculum where the model efficiently learns from semantically informative samples while avoiding overfitting to redundant or noisy signals.

4 Experiments

4.1 Experimental Setup

Training Datasets. To validate SDAE's robustness across different training settings, we conduct experiments on two representative datasets. **MATH** [28] contains 7,500 complex competition mathematics problems spanning algebra, geometry, number theory, and other domains. Following prior work [12], we use the training set for reinforcement learning. **DAPO-Math-17K** [22] is a dataset containing 17,000 mathematics problems with integer answers, specifically designed for large-scale LLM reinforcement learning and carefully curated to ensure accurate reward signals. In accordance with Cheng et al. [10], we adopt the same prompt templates and verifiers for experiments.

Model and Baselines. Following prior work [12], we use **Qwen2.5-Math-7B** [29] as the primary backbone for our main experiments. We compare SDAE against mainstream RL algorithms and recent exploration-enhanced methods, including GRPO, Entropy Adv, AEnt, PSR, NSR, and W-REINFORCE. In addition, to evaluate whether SDAE generalizes beyond a single Qwen-based 7B model, we further report results in Section 4.3 on **Llama-3.1-8B-Instruct** [30], representing a different model family, and on **Qwen3-4B** (non-thinking mode) [21], representing a different model scale. PSR and NSR selectively update the policy model using only correct or incorrect responses, respectively. Entropy Adv encourages longer reasoning chains by adding token-level entropy rewards $\phi(H)$. AEnt is a recent entropy-control baseline. W-REINFORCE is a REINFORCE variant that reduces the weight of positive sample reinforcement while maintaining negative sample reinforcement, balancing accuracy and diversity. More details are provided in Appendix E.

RL Training Configuration. We train models using the VERL framework [31]. For both datasets, we use the prompt templates shown in Appendix F and implementation details in Appendix G to further elaborate the training setup.

Evaluation Benchmarks and Metrics. We evaluate on AIME 2024/2025 [32], AMC23 [33], and MATH/MATH500 [28]. For Qwen2.5-Math-7B and Llama-3.1-8B-Instruct, we use a sampling temperature of 0.6, a maximum response length of 4K tokens, and top-p sampling with $p = 0.95$, and report Pass@k with $k \in \{1, 2, 4, 8, 16, 32, 64, 128, 256\}$. For Qwen3-4B, we use a temperature of 0.7, top-p of 0.8, and top-k of 20, and report Pass@k with $k \in \{1, 2, 4, 8, 16, 32, 64\}$. Following Zhu et al. [12], we use Pass@k to evaluate reasoning capability boundaries, with details provided in Appendix I.

4.2 Main Results

As shown in Table 1, we report Pass@k and Pass@1 results for Qwen2.5-Math-7B trained on DAPO-Math-17K. GRPO w/ SDAE consistently outperforms baseline methods across all benchmarks. On AIME25, SDAE achieves Pass@256 of 63.3% and Pass@1 of 20.0%, representing improvements of +5.9% and +3.7% over vanilla GRPO, respectively. Compared to GRPO w/ Entropy Advantage, SDAE's Pass@1 is +2.4% higher, demonstrating that semantic-level diversity provides more effective exploration signals than token-level entropy. Notably, on AMC23, SDAE achieves a +12.8% Pass@1 improvement, substantially surpassing entropy-based methods (+2.9%). SDAE also matches or exceeds multiple strong baselines across benchmarks, including Qwen2.5-Math-Instruct, Eurus-2-PRIME, Oat-Zero, and GPG, validating its effectiveness as a general exploration enhancement technique for LLM reinforcement learning.

Table 1: **Pass@k and Pass@1 results on MATH500, AIME 24/25 and AMC23.** †: results from Chu et al. [34], ‡: results from Cheng et al. [10]. **Green** type denotes the performance gain of SDAE (blue background) over related baselines (gray background).

Method	Venue	AIME25		AIME24		AMC23		MATH500	
		Pass@256	Pass@1	Pass@256	Pass@1	Pass@128	Pass@1	Pass@16	Pass@1
Qwen2.5-Math	Qwen	53.3	6.2	70.0	14.2	90.0	40.3	90.6	57.2
Qwen2.5-Math-Instruct	Qwen	53.3	8.7	60.0	11.3	92.4	54.7	93.5	79.8
Eurus-2-PRIME†	arXiv'25	-	-	-	26.7	-	57.8	-	79.2
Oat-Zero†	COLM'25	-	-	-	30.0	-	55.4	-	80.6
GPG†	ICLR'26	-	-	-	33.3	-	65.0	-	80.0
+ GRPO‡	-	57.4	16.3	83.3	30.9	92.8	66.9	94.6	83.0
+ GRPO w/ Entropy Adv‡	AAAI'26	63.6	17.6	80.0	33.7	95.2	69.8	94.8	83.1
+ GRPO w/ SDAE	**Ours**	63.3	20.0	80.0	32.9	95.6	79.7	95.4	85.2
△ Gain		**+5.9**	**+3.7**	**-3.3**	**+2.0**	**+2.8**	**+12.8**	**+0.8**	**+2.2**

Table 2 presents Pass@k results trained on the full MATH dataset. REINFORCE w/ SDAE achieves the best results across nearly all k values on MATH (Pass@1 81.0%, Pass@256 99.2%) and AIME25 (Pass@1 13.2%, Pass@256 66.7%). Both SDAE variants achieve 100.0% Pass@256 on AMC23, while multiple baselines fail to maintain this exploration capability. Unlike PSR, which improves Pass@1 but degrades high-k performance, SDAE improves across the entire Pass@k spectrum. This advantage also holds when compared with AEnt, an entropy-control baseline: under the same Qwen2.5-Math-7B backbone and MATH training setup, GRPO w/ SDAE consistently outperforms AEnt on AIME25 across all Pass@k metrics, and on AMC23 from Pass@4 onward, while AEnt is slightly stronger only at very low k (Pass@1 and Pass@2). On AIME25, SDAE not only surpasses existing RL baselines but also breaks through the base model's Pass@256 ceiling, demonstrating its ability to transcend the base model's inherent reasoning boundaries. Overall, these results suggest that SDAE provides more effective exploration signals than entropy control alone, especially when reasoning performance under larger decoding budgets is of primary interest.

4.3 Generalization Across Model Families and Scales

To test whether SDAE extends beyond the single Qwen2.5-Math-7B setting, we further evaluate it on a different model family, Llama-3.1-8B-Instruct, and on a different parameter scale, Qwen3-4B. All models are trained on MATH, and we report Pass@k on MATH, AIME25, and AMC23.

As shown in Figure 3, on Llama-3.1-8B-Instruct, unlike the Qwen-based backbones, applying RL generally degrades inference-scaling performance relative to the base model. Although Pass@1 improves after RL training on MATH and AMC23, high-k performance remains substantially below the base model. For example, SDAE reaches Pass@256 of 90.1 / 33.3 / 92.5 on MATH / AIME25 / AMC23, whereas the corresponding base-model results are 93.4 / 40.0 / 100.0. This contrast suggests that the backbone itself often determines whether RL can reliably bring gains, consistent with recent observations in the literature [12, 35, 36, 6]. Among all RL variants, however, SDAE preserves the strongest scaling behavior: it achieves the best high-k results on all three benchmarks and therefore causes the least damage to the base model's reasoning ceiling.

Figure 4 further shows that, on Qwen3-4B, SDAE also remains effective under a change in model scale. Relative to vanilla GRPO, SDAE improves MATH Pass@1 / Pass@64 from 93.9 / 98.2 to 94.7 / 98.6; on AIME25, it improves Pass@1 / Pass@16 from 56.8 / 76.4 to 62.7 / 80.3; and on AMC23, it raises Pass@4 from 98.1 to 99.9 and reaches 100.0 once $k \geq 8$. Although NSR remains competitive

Table 2: **Pass@k results on MATH, AIME25 and AMC23 with Qwen2.5-Math-7B. Bold** and underlined types denote the best and second-best results for each k. $\star$: results from Zhu et al. [12].

Method	**Pass@k**								
k	1	2	4	8	16	32	64	128	256
MATH									
Base Model*	63.2	76.0	83.7	88.4	91.6	93.7	95.2	96.2	96.9
GRPO*	76.3	81.7	85.6	88.4	90.6	92.3	93.6	94.7	95.5
PPO*	76.6	82.6	86.7	89.6	91.7	93.4	94.7	95.6	96.3
PSR*	74.1	78.3	81.6	84.1	86.2	87.9	89.3	90.4	91.2
NSR*	75.7	82.4	86.9	90.1	92.4	94.1	95.3	96.2	96.9
W-REINFORCE*	76.6	82.8	87.1	90.2	92.4	94.1	95.3	96.1	96.7
AEnt	77.0	82.5	87.0	89.4	91.4	92.7	93.9	94.9	95.6
REINFORCE w/ SDAE	**81.0**	**85.0**	**89.3**	**93.2**	**97.8**	**98.3**	**98.7**	**98.9**	**99.2**
GRPO w/ SDAE	77.1	83.5	87.1	90.7	92.1	95.0	95.9	96.6	97.2
AIME25									
Base Model*	6.1	9.7	13.8	17.9	22.2	26.5	30.8	36.6	46.7
GRPO*	10.3	14.7	19.4	24.0	28.4	32.8	37.3	42.5	50.0
PPO*	8.5	13.2	18.0	22.5	26.6	30.3	33.8	37.9	43.3
PSR*	11.6	14.1	16.2	18.6	21.7	25.7	30.9	36.9	43.3
NSR*	10.0	14.6	19.2	24.1	29.3	34.6	40.2	46.0	53.3
W-REINFORCE*	10.6	15.3	20.0	24.7	29.7	34.6	40.5	47.8	56.7
AEnt	10.7	15.0	19.3	23.1	26.2	29.0	32.0	35.6	40.0
REINFORCE w/ SDAE	**13.2**	**18.0**	**22.4**	**26.9**	**31.9**	**37.9**	**45.2**	**54.6**	**66.7**
GRPO w/ SDAE	11.7	15.8	20.0	24.5	30.0	36.3	44.1	53.6	63.3
AMC23									
Base Model*	41.0	56.2	69.2	78.9	85.1	89.1	92.9	97.2	**100.0**
GRPO*	61.7	68.7	74.6	80.0	85.1	89.7	93.4	95.9	97.5
PPO*	62.0	70.0	76.1	80.9	85.3	89.5	93.1	96.0	97.5
PSR*	62.6	69.9	74.5	77.5	80.3	83.5	87.2	90.6	92.5
NSR*	60.9	70.0	77.4	83.2	87.6	91.1	94.5	97.9	**100.0**
W-REINFORCE*	62.0	70.0	77.0	83.1	87.8	91.8	95.2	97.1	97.5
AEnt	**63.7**	**70.7**	75.8	80.1	84.3	88.6	93.0	96.5	97.5
REINFORCE w/ SDAE	61.4	70.1	77.0	82.8	87.6	92.2	**96.4**	**99.3**	**100.0**
GRPO w/ SDAE	62.5	70.6	**77.5**	**83.5**	**88.3**	**92.6**	96.1	98.9	**100.0**

on a few high-k metrics, SDAE overall shows a more stable generalization pattern across model scales.

Overall, these results indicate that SDAE is not tied to a particular backbone or parameter scale. Under cross-scale evaluation, it delivers stable performance gains; under cross-family evaluation, even when the backbone itself limits the overall benefit of RL relative to the base model, SDAE still preserves the strongest inference-scaling behavior among RL variants. This supports its effectiveness and robustness as a general exploration-enhancement mechanism.

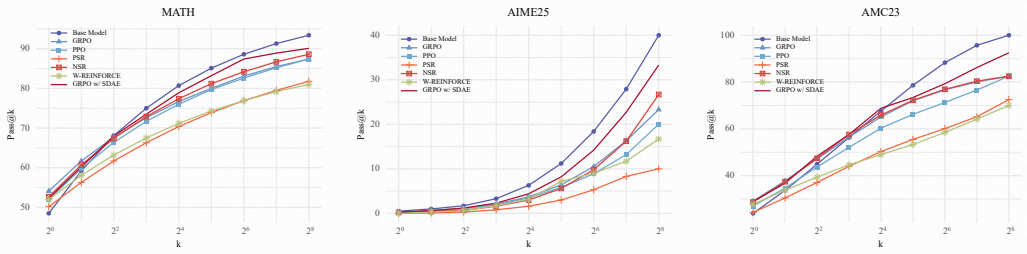


Figure 3: **Pass@k performance of Llama-3.1-8B-Instruct on MATH** using different RL algorithms across various benchmarks.

4.4 Analysis of Exploration Capability via Pass@k

To further analyze the effect of SDAE on exploration, we compare different methods across the full Pass@k spectrum (Figure 5). The results show that SDAE improves not only Pass@1, but also better preserves and enhances the high-k reasoning ceiling, indicating a superior balance between exploration and exploitation. Detailed discussion is deferred to Appendix J.

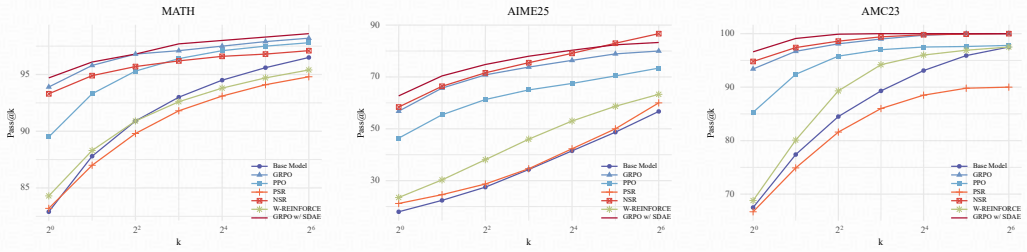


Figure 4: **Pass@k performance of Qwen3-4B (Non-Thinking) on MATH** using different RL algorithms across various benchmarks.

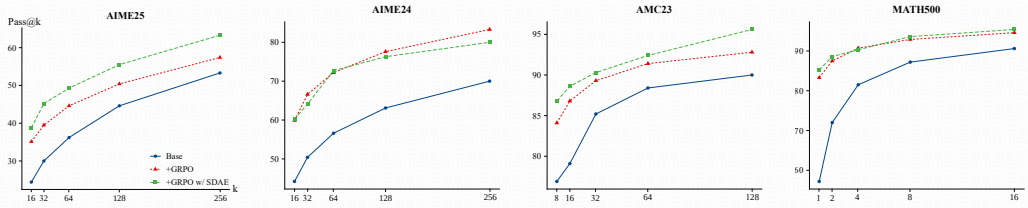


Figure 5: **Pass@k performance of Qwen2.5-Math-7B on DAPO-Math-17K** using different RL algorithms across various benchmarks.

4.5 Entropy Dynamics: Mitigating Strategy Collapse

We further analyze how SDAE mitigates policy collapse through the evolution of policy entropy during training. As shown in Figure 6, compared with vanilla GRPO, SDAE maintains a substantially higher and more stable entropy level, suggesting that it better preserves exploration capacity; its effect is also comparable to existing entropy-control methods. Detailed analysis is provided in Appendix K.

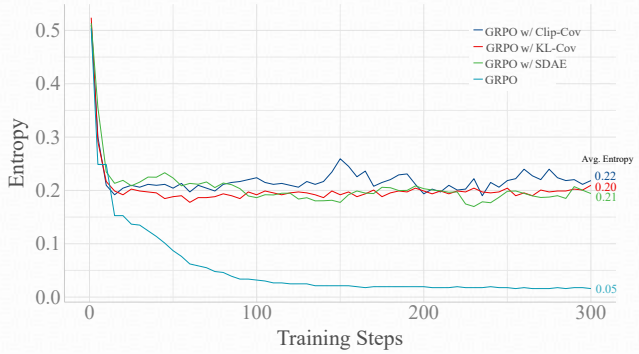


Figure 6: Policy entropy dynamics on the DAPO-Math-17K dataset. SDAE maintains a stable entropy level (0.21) throughout training, effectively mitigating the catastrophic entropy collapse seen in vanilla GRPO (0.05). Its performance is comparable to gradient-level regularization methods like Clip-Cov and KL-Cov.

4.6 Analysis of Training Dynamics and Behavior Evolution

From the perspective of training dynamics, as shown in Figure 7, SDAE exhibits healthier behavior evolution in terms of accuracy, policy entropy, and the efficiency of discovering correct solution paths. Compared with standard RL objectives, SDAE avoids premature pure-exploitation convergence, while more steadily improving greedy decoding accuracy and more effectively uncovering correct reasoning paths. A fuller interpretation of these trajectories and behaviors is deferred to Appendix L.

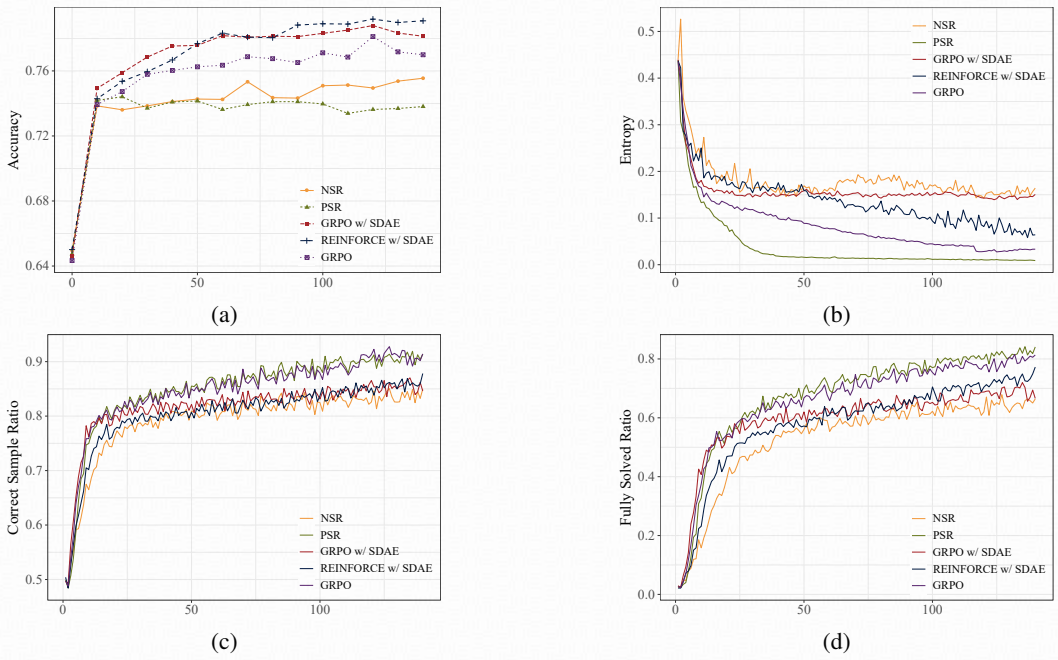


Figure 7: Training dynamics on MATH dataset under GRPO, PSR, NSR, GRPO w/ SDAE and REINFORCE w/ SDAE across training steps, including (a) greedy decoding accuracy on the math test set, (b) the model's entropy, (c) the ratio of correct responses per batch on the training set, and (d) the proportion of fully-solved prompts per batch (i.e., all rollouts are correct) on the training set.

4.7 Impact of Embedding Spaces on Semantic Exploration

We also investigate the effect of different semantic embedding spaces on SDAE. The results show that the default LSA variant is overall the strongest or most stable choice, while larger external encoders consistently improve ESE performance, suggesting that the gains of SDAE mainly come from semantic-level diversity modeling itself rather than reliance on any particular external encoder. Detailed comparisons are provided in Appendix M.

4.8 Analyses of Semantic Diversity and Reasoning Quality

We further analyze the effect of SDAE on the semantic diversity and process quality of reasoning. The results show that its gains are not limited to final-answer accuracy: on AIME25, SDAE yields more semantically distinct correct solution strategies and higher step-level reasoning accuracy at comparable reasoning lengths; qualitative case studies further show that it is less prone to collapsing to a single successful template. Detailed quantitative and qualitative evidence is provided in Appendix N.

5 Conclusion

This paper proposes Semantic-Diversity-Aware Exploration (SDAE), a novel method that guides LLM reasoning exploration through semantic space geometry. Our core insight is that binary reward signals discard the semantic geometric information of the response space, while correctness and semantic distance capture complementary, non-redundant information about a response's exploration value. SDAE differentially modulates advantage functions by computing responses' semantic distance to the positive centroid, adaptively balancing exploration and exploitation without additional hyperparameters. Experiments demonstrate that SDAE outperforms strong baselines across multiple benchmarks, achieving a 13.3% improvement in Pass@256 on AIME25, effectively breaking through the reasoning ceiling of the baseline model. Our work demonstrates that the geometric structure of semantic space can precisely capture the nuanced exploration value of responses, providing a new perspective for balancing exploration and exploitation.

Acknowledgments

This work was financially supported by the National Natural Science Foundation of China (62437002, 62607078, 62307015, 62293554), Hubei Provincial Natural Science Foundation of China (2026AFB706, 2023AFA020) and Natural Science Foundation of Wuhan (2025040601020160).

References

[1] D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi *et al.*, "Deepseek-r1 incentivizes reasoning in llms through reinforcement learning," *Nature*, vol. 645, no. 8081, pp. 633–638, 2025.

[2] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu *et al.*, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

[3] Y. Wang, Q. Yang, Z. Zeng, L. Ren, L. Liu, B. Peng, H. Cheng, X. He, K. Wang, J. Gao *et al.*, "Reinforcement learning for reasoning in large language models with one training example," *arXiv preprint arXiv:2504.20571*, 2025.

[4] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[5] X. Dang, C. Baek, K. Wen, Z. Kolter, and A. Raghunathan, "Weight ensembling improves reasoning in language models," *arXiv preprint arXiv:2504.10478*, 2025.

[6] Y. Yue, Z. Chen, R. Lu, A. Zhao, Z. Wang, S. Song, and G. Huang, "Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?" *arXiv preprint arXiv:2504.13837*, 2025.

[7] G. Cui, Y. Zhang, J. Chen, L. Yuan, Z. Wang, Y. Zuo, H. Li, Y. Fan, H. Chen, W. Chen *et al.*, "The entropy mechanism of reinforcement learning for reasoning language models," *arXiv preprint arXiv:2505.22617*, 2025.

[8] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*. Pmlr, 2018, pp. 1861–1870.

[9] B. D. Ziebart, A. L. Maas, J. A. Bagnell, A. K. Dey *et al.*, "Maximum entropy inverse reinforcement learning." in *Aaai*, vol. 8. Chicago, IL, USA, 2008, pp. 1433–1438.

[10] D. Cheng, S. Huang, X. Zhu, B. Dai, W. X. Zhao, Z. Zhang, and F. Wei, "Reasoning with exploration: An entropy perspective," *arXiv preprint arXiv:2506.14758*, 2025.

[11] S. Wang, L. Yu, C. Gao, C. Zheng, S. Liu, R. Lu, K. Dang, X. Chen, J. Yang, Z. Zhang *et al.*, "Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning," *arXiv preprint arXiv:2506.01939*, 2025.

[12] X. Zhu, M. Xia, Z. Wei, W.-L. Chen, D. Chen, and Y. Meng, "The surprising effectiveness of negative reinforcement in llm reasoning," *arXiv preprint arXiv:2506.01347*, 2025.

[13] X. Tang, Y. Zhan, Z. Li, W. X. Zhao, Z. Zhang, Z. Wen, Z. Zhang, and J. Zhou, "Rethinking sample polarity in reinforcement learning with verifiable rewards," *arXiv preprint arXiv:2512.21625*, 2025.

[14] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *Advances in neural information processing systems*, vol. 30, 2017.

[15] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[16] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Advances in neural information processing systems*, vol. 36, pp. 53 728–53 741, 2023.

[17] N. Lambert, J. Morrison, V. Pyatkin, S. Huang, H. Ivison, F. Brahman, L. J. V. Miranda, A. Liu, N. Dziri, S. Lyu *et al.*, "Tulu 3: Pushing frontiers in open language model post-training," *arXiv preprint arXiv:2411.15124*, 2024.

[18] A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney *et al.*, "Openai o1 system card," *arXiv preprint arXiv:2412.16720*, 2024.

[19] Q. Team, "Qwq-32b: Embracing the power of reinforcement learning," 2025.

[20] K. Team, A. Du, B. Gao, B. Xing, C. Jiang, C. Chen, C. Li, C. Xiao, C. Du, C. Liao *et al.*, "Kimi k1. 5: Scaling reinforcement learning with llms," *arXiv preprint arXiv:2501.12599*, 2025.

[21] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[22] Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, W. Dai, T. Fan, G. Liu, L. Liu *et al.*, "Dapo: An open-source llm reinforcement learning system at scale," *arXiv preprint arXiv:2503.14476*, 2025.

[23] T. Li, Y. Zhang, P. Yu, S. Saha, D. Khashabi, J. Weston, J. Lanchantin, and T. Wang, "Jointly reinforcing diversity and quality in language model generations," *arXiv preprint arXiv:2509.02534*, 2025.

[24] R. Jin, P. Gao, Y. Ren, Z. Han, T. Zhang, W. Huang, W. Liu, J. Luan, and D. Xiong, "Revisiting entropy in reinforcement learning for large reasoning models," *arXiv preprint arXiv:2511.05993*, 2025.

[25] Z. Liang, S. Lu, W. Yu, K. Panaganti, Y. Zhou, H. Mi, and D. Yu, "Can llms guide their own exploration? gradient-guided reinforcement learning for llm reasoning," *arXiv preprint arXiv:2512.15687*, 2025.

[26] Y. Zhou, Z. Liang, H. Liu, W. Yu, K. Panaganti, L. Song, D. Yu, X. Zhang, H. Mi, and D. Yu, "Evolving language models without labels: Majority drives selection, novelty promotes variation," *arXiv preprint arXiv:2509.15194*, 2025.

[27] M. Chen, G. Chen, W. Wang, and Y. Yang, "Seed-grpo: Semantic entropy enhanced grpo for uncertainty-aware policy optimization," *arXiv preprint arXiv:2505.12346*, 2025.

[28] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt, "Measuring mathematical problem solving with the math dataset," *arXiv preprint arXiv:2103.03874*, 2021.

[29] A. Yang, B. Zhang, B. Hui, B. Gao, B. Yu, C. Li, D. Liu, J. Tu, J. Zhou, J. Lin *et al.*, "Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement," *arXiv preprint arXiv:2409.12122*, 2024.

[30] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[31] G. Sheng, C. Zhang, Z. Ye, X. Wu, W. Zhang, R. Zhang, Y. Peng, H. Lin, and C. Wu, "Hybridflow: A flexible and efficient rlhf framework," in *Proceedings of the Twentieth European Conference on Computer Systems*, 2025, pp. 1279–1297.

[32] MAA, "American invitational mathematics examination (aime)," https://maa.org/math-competitions/aime, 2025, accessed: 2026-01-07.

[33] ——, "American mathematics competitions (amc)," https://maa.org/math-competitions/amc, 2023, accessed: 2026-01-07.

[34] X. Chu, H. Huang, X. Zhang, F. Wei, and Y. Wang, "Gpg: A simple and strong reinforcement learning baseline for model reasoning," *arXiv preprint arXiv:2504.02546*, 2025.

[35] Z. Wang, F. Zhou, X. Li, and P. Liu, "Octothinker: Mid-training incentivizes reinforcement learning scaling," *arXiv preprint arXiv:2506.20512*, 2025.

[36] E. Yeo, Y. Tong, M. Niu, G. Neubig, and X. Yue, "Demystifying long chain-of-thought reasoning in llms," *arXiv preprint arXiv:2502.03373*, 2025.

[37] G. Cui, L. Yuan, Z. Wang, H. Wang, Y. Zhang, J. Chen, W. Li, B. He, Y. Fan, T. Yu *et al.*, "Process reinforcement through implicit rewards," *arXiv preprint arXiv:2502.01456*, 2025.

[38] Z. Liu, C. Chen, W. Li, P. Qi, T. Pang, C. Du, W. S. Lee, and M. Lin, "Understanding r1-zero-like training: A critical perspective," *arXiv preprint arXiv:2503.20783*, 2025.

[39] H. Shen, "On entropy control in llm-rl algorithms," *arXiv preprint arXiv:2509.03493*, 2025.

[40] Hugging Face, "Math-verify: A robust mathematical expression evaluation system," https://github.com/huggingface/Math-Verify, 2025, accessed: 2026-01-07.

[41] M. Chen, "Evaluating large language models trained on code," *arXiv preprint arXiv:2107.03374*, 2021.

A Theoretical Connection to Gradient-Guided Exploration

We provide additional context on the theoretical relationship between our Latent Semantic Aggregation (LSA) approach and recent work on gradient-guided exploration.

G^{2}RL [25] analyzes the source of exploration signals from the perspective of first-order update geometry, demonstrating that external semantic embeddings are inherently flawed as proxies for exploration. This provides theoretical justification for using the policy model's internal representations. Specifically, G²RL proves that for autoregressive language models, parameter gradients at any layer can be decomposed as:

$$\nabla_{\theta_k} \ell(x, y) = \frac{1}{T} \mathcal{L}_k(x, y) \Phi(x, y), \tag{10}$$

where $\Phi(x, y) = \sum_t W(e(y_t) - p_t)$ is a sequence-level gradient sensitivity feature derived from the last-layer, and $\mathcal{L}_k(x, y)$ is a trajectory-dependent linear operator. This factorization theorem reveals that the last-layer serves as the unique first-order sensitivity bottleneck through which all trajectory-specific optimization information must pass.

It is worth noting that while G^{2}RL's sequence feature $\Phi(x, y)$ represents an exact gradient sensitivity measure, our LSA employs a simplified approximation using mean-pooled hidden states: $v_i = \text{MeanPool}(\{h_{i,t}\}_{t=1}^{T_i})$. Despite this difference in mathematical formulation, the core insight remains consistent: the policy model's internal representation space reflects optimization-relevant geometric structure more faithfully than external encoders. Our ablation study in Table 8 empirically validates the effectiveness of this design choice.

Furthermore, while G^{2}RL guides exploration directly in the gradient feature space, our SDAE operates in the semantic embedding space, achieving finer-grained advantage modulation through the gravitational mechanism. This distinction allows SDAE to focus on the semantic diversity of reasoning paths rather than raw gradient orthogonality, which we find to be more effective for incentivizing novel solution strategies while suppressing divergent errors.

B Handling Homogeneous Rollouts (All-Correct or Zero-Correct Cases)

In the practical implementation of SDAE, we encounter cases where all N rollouts for a single prompt q are either entirely correct ($r_i = 1, \forall i$) or entirely incorrect ($r_i = 0, \forall i$). We handle these homogeneous groups differently depending on the underlying RL algorithm:

B.1 Integration with GRPO

For GRPO w/ SDAE, we strictly follow the advantage computation logic of vanilla GRPO. The standard advantage A_i is defined by normalizing the rewards within the sampled group:

$$A_i = \frac{R(o_i) - \text{mean}(\{R(o)\})}{\text{std}(\{R(o)\})}. \tag{11}$$

When all responses share the same reward label (e.g., all correct or all incorrect), the standard deviation becomes zero (or the numerator vanishes), resulting in $A_i = 0$ for all i. Consequently, the semantically-aware advantage also becomes zero:

$$A_i^{SDAE} = A_i \cdot \hat{SG}_i = 0. \tag{12}$$

As a result, these samples provide no gradient signal and do not trigger policy updates. This design ensures that SDAE maintains consistency with the baseline GRPO mechanism without introducing unintended optimization biases.

B.2 Integration with REINFORCE

In REINFORCE w/ SDAE, the calculation of the gravitational source μ^+ typically requires the set of correct responses $\mathcal{Y}^+$. We handle the edge cases as follows:

- **All-Correct Case:** If all rollouts are correct ($\mathcal{Y}^+ = \mathcal{Y}$), the centroid μ^+ is calculated as the mean-pool of all N sampled hidden states v_i.
- **Zero-Correct (All-Wrong) Case:** When no correct response is found ($\mathcal{Y}^+ = \emptyset$), the standard definition of μ^+ becomes undefined. In this scenario, we adopt a consensus-based fallback, using the centroid of all sampled rollouts as the proxy gravitational source:

$$\mu^+ = \frac{1}{N} \sum_{i=1}^{N} v_i \quad \text{if} \quad \mathcal{Y}^+ = \emptyset. \tag{13}$$

  This allows the model to still distinguish between Local Errors (semantically near the group's consensus) and Divergent Errors. By attenuating penalties for local errors relative to the group consensus, we preserve the model's potential for iterative refinement even when no correct path is discovered in the current batch.

C Normalization of Semantic Gravity

In our implementation, the Semantic Gravity (SG) is not mapped to a fixed range via hard clipping. Instead, we employ a batch-level mean normalization strategy to modulate the advantages. Specifically, for a set of sampled responses for a prompt, the raw semantic distances are scaled by their batch mean:

$$\hat{SG}_i = \frac{SG_i}{\frac{1}{N}\sum_{j=1}^{N} SG_j + \epsilon}, \tag{14}$$

where SG_i represents the raw cosine distance between the response embedding and the positive centroid. This normalization provides several key benefits:

- **Scale Invariance:** By ensuring the average multiplier is approximately 1.0, we preserve the overall magnitude of the gradients, preventing the advantage values from vanishing or exploding as the model's embedding space shifts.
- **Relative Novelty:** This approach emphasizes the relative semantic novelty of a response within its current sampling group rather than its absolute position in the embedding space.
- **Stability:** It avoids the information loss associated with hard clipping (e.g., mapping all distant samples to 1.0), allowing the model to distinguish between "moderately novel" and "highly divergent" paths.

D Analysis of Computational and Memory Overhead

The overhead introduced by Latent Semantic Aggregation (LSA) is negligible compared to the standard training and inference costs of LLMs. We analyze this from two perspectives:

- **Computational Efficiency:**
  - **Intrinsic Reuse:** LSA directly utilizes the last-layer hidden states already generated during the policy model's forward pass. Since these states are required for generating logits, LSA incurs zero additional inference cost.
  - **Low Complexity:** The mean pooling operation $v_i = \text{MeanPool}(\{h_{i,t}\})$ has a linear time complexity of $O(T \times d)$ (where T is sequence length and d is hidden dimension). This is several orders of magnitude smaller than the $O(T^2 \times d)$ complexity of the self-attention mechanism.

  - **Self-Contained Advantage:** Unlike methods requiring External Semantic Encoders (ESE), SDAE avoids the massive computational burden of running an additional 0.6B or 4B parameter model for every rollout.
- **Memory Efficiency:**
  - **Vector Compression:** While storing per-token hidden states for a full batch might seem demanding, SDAE only requires the pooled semantic vector $v_i \in \mathbb{R}^d$ for advantage modulation.
  - **Immediate Release:** The high-dimensional per-token tensors can be discarded immediately after the mean-pooling step during the rollout phase. The resulting storage overhead is merely $N \times d$ per prompt (where N is the number of rollouts), which is marginal compared to the memory occupied by model weights and KV caches.

E Detailed Description of Baselines

In this section, we provide detailed descriptions of baseline methods used for comparison in our experiments. Specifically, we compare our method against GRPO [2], PRIME [37], Oat-Zero [38], GPG [34], polarity-level advantage shaping methods (i.e., PSR, NSR, and W-REINFORCE [12]), and token-level advantage shaping methods (i.e., w/ Entropy Advantage [10]).

- **GRPO** [2] was proposed in DeepSeekMath as a simplified alternative to PPO, eliminating the need for a separate critic model. Instead of learning a value function, GRPO uses the mean reward within a sampled group as the baseline.

  Objective function:

  The GRPO objective maximizes the following clipped surrogate objective:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{[q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(O|q)]} \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}\hat{A}_{i,t}, \text{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}, 1-\epsilon, 1+\epsilon\right)\hat{A}_{i,t}\right) - \beta\mathbb{D}_{KL}[\pi_\theta||\pi_{ref}]\right\}, \tag{15}$$

  here, ε is the clipping hyperparameter, β is the KL penalty coefficient, and π_{ref} denotes the reference policy.

  Advantage computation:

  Unlike PPO which learns a value function, GRPO computes advantages using group-level statistics:

$$\hat{A}_i = \frac{R(o_i) - \text{mean}(\{R(o_1), \ldots, R(o_G)\})}{\text{std}(\{R(o_1), \ldots, R(o_G)\})}, \tag{16}$$

  where G outputs are sampled for each question and scored by a reward model.
- **PRIME** [37] introduces implicit process rewards, providing dense token-level feedback during RL training:

  Core mechanism:

  The implicit PRM computes process rewards as:

$$r_\phi(y_t) = v_\phi(y_{<t+1}) - v_\phi(y_{<t}), \tag{17}$$

  where $v_\phi(y_{<t}) = \sum_{i=1}^{t-1}\beta\log\frac{\pi_\phi(y_i|y_{<i})}{\pi_{ref}(y_i|y_{<i})}$ is the implicit value function.
- **Oat-Zero** [38] demonstrates that RL can be directly applied to base models without SFT warm-up, while critically examining optimization biases in GRPO.
- **GPG** [34] proposes a simplified RL approach that directly optimizes the original objective without using surrogate losses:

  Loss function:

$$\mathcal{L}_{\text{GPG}} = -\log\pi_\theta(o)\cdot A, \tag{18}$$

  where $A = \alpha\cdot(R(o) - \text{mean}(\{R(o)\}))$

- **PSR** and **NSR** are polarity-level advantage shaping methods proposed by Zhu et al. [12], selectively updating policies based on response correctness.
- **W-REINFORCE** [12] is a REINFORCE variant that reduces the weight of the positive sample reinforcement component while maintaining negative sample reinforcement, balancing accuracy and diversity.

  Formulation:

  The method reduces weights on positive advantage terms while maintaining full weights on negative advantages:

$$\mathcal{L}_{\text{W-REINFORCE}} = -\mathbb{E}\left[w \cdot \mathbb{1}[A > 0] \cdot A \cdot \log \pi_\theta(o) + \mathbb{1}[A < 0] \cdot A \cdot \log \pi_\theta(o)\right], \tag{19}$$

  where $w < 1$ is the weight factor for positive samples.
- **Entropy Advantage** [10] proposes encouraging exploratory reasoning behaviors by adding entropy-based terms to standard RL advantages.

  Entropy-based advantage shaping:

  For each token o_t, the entropy of the current policy over vocabulary $\mathcal{V}$ is:

$$\mathcal{H}_t = -\sum_{v \in \mathcal{V}} \pi_\theta(v|q, o_{<t}) \log \pi_\theta(v|q, o_{<t}). \tag{20}$$

  The shaped advantage is:

$$A_t^{\text{shaped}} = A_t + \psi(\mathcal{H}_t), \tag{21}$$

  where:

$$\psi(\mathcal{H}_t) = \min\left(\alpha \cdot \mathcal{H}_t^{\text{detach}}, \frac{|A_t|}{\kappa}\right), \quad \alpha > 0, \kappa > 1$$

- **AEnt** [39] is a recent entropy-control method for LLM-RL based on adaptive clamped entropy regularization. It replaces conventional entropy regularization with a clamped entropy bonus defined on a reduced token space and automatically adjusts the entropy coefficient during training. In our comparison, we use AEnt as a recent entropy-control baseline to test whether SDAE provides benefits beyond entropy-based exploration control alone.

F Prompt Templates and Reward Function

We adopt model-specific prompt templates for Qwen2.5-Math-7B, Llama-3.1-8B-Instruct, and Qwen3-4B (non-thinking mode), as shown in Tables Table 3-5, and use Math-Verify [40] as the verifier. For Qwen2.5-Math-7B and Llama-3.1-8B-Instruct, the instruction content is identical and only the model-specific chat-template syntax differs. For Qwen3-4B, we additionally insert an empty <think> ... </think> block to match its non-thinking format.

Table 3: Prompt template for Qwen2.5-Math-7B.

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
{input}
Please reason step by step, and put your final answer within
\boxed{}.<|im_end|>
<|im_start|>assistant
```

Table 4: Prompt template for Llama-3.1-8B-Instruct.

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>
You are a helpful
assistant.<|eot_id|><|start_header_id|>user<|end_header_id|>
{input}
Please reason step by step, and put your final answer within
\boxed{}.<|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

Table 5: Prompt template for Qwen3-4B.

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
{input}
Please reason step by step, and put your final answer within
\boxed{}.<|im_end|>
<|im_start|>assistant
<think>

</think>
```

G RL Training Configuration

For MATH, we use hyperparameters shown in Table 6, with a prompt batch size of 1,024 and 8 rollouts per prompt. The sampling temperature during training is set to 1.0, and maximum context length is set to 4,096 tokens. We update the model with a mini-batch size of 256 and learning rate of 1e-6. For DAPO-Math-17K, we use hyperparameters shown in Table 7, with a prompt batch size of 512 and 8 rollouts per prompt. The sampling temperature during training is set to 1.0, and maximum context length is set to 4,096 tokens. We update the model with a mini-batch size of 512 and learning rate of 1e-6.

Table 6: Hyperparameters for MATH

Hyperparameter	**Value**
Optimizer	AdamW
Learning rate	$1e^{-6}$
Batch size	1024
Samples per prompt	8
Mini-batch size	256
Max context length	4K
Temperature	1.0

Table 7: Hyperparameters for DAPO-Math-17K

Hyperparameter	**Value**
Optimizer	AdamW
Learning rate	$1e^{-6}$
Batch size	512
Samples per prompt	8
Mini-batch size	512
Max context length	4K
Temperature	1.0

H Detailed Description of Evaluation Benchmarks

In this section, we provide detailed descriptions of evaluation benchmarks used in our experiments. Our evaluation focuses on competition-level mathematical reasoning tasks that require multi-step logical reasoning and complex problem-solving abilities.

- **MATH / MATH500** [28] is a comprehensive benchmark designed to measure machine learning models' mathematical problem-solving abilities. It contains 12,500 challenging competition mathematics problems sourced from high school mathematics competitions such as AMC 10, AMC 12, AIME, and other renowned mathematical olympiads.

Dataset composition:

- **Total problems:** 12,500 (7,500 training + 5,000 test)
- **MATH500:** A curated subset of 500 representative test problems, commonly used for efficient evaluation

- **AIME** [32] is a prestigious examination following the AMC 10 and AMC 12 competitions, containing 15 problems with a 3-hour time limit. It serves as a qualifier for the USA Mathematical Olympiad (USAMO) and represents one of the most challenging standardized mathematical reasoning benchmarks.

  Versions used:

  - **AIME 2024:** 30 problems (2024 AIME I + AIME II)
  - **AIME 2025:** 30 problems (2025 AIME I + AIME II)

- **AMC** [33] is a series of mathematics competitions organized by the Mathematical Association of America (MAA). We specifically use AMC23 (40 problems), which includes problems from AMC 12.

I Detailed Formulation of Pass@k Metric

When evaluating generative model performance, directly sampling k results and checking for correct answers leads to high variance. To obtain more robust, unbiased evaluation results, we adopt the combinatorics-based expected estimation method proposed by Chen et al. [41]. It generates n samples ($n \geq k$) for each problem, counts the number of correct responses c, and computes the unbiased estimate of Pass@k as:

$$Pass@k = \mathbb{E}_{x\sim\mathcal{D}}\left[1 - \frac{\binom{n-c}{k}}{\binom{n}{k}}\right]. \tag{22}$$

J Analysis of Exploration Capability via Pass@k

To better understand how SDAE affects model exploration capability, we analyze Pass@k curves across different k values. As shown in Figure 5 (on DAPO-Math-17K) and Table 2 (on MATH), we compare the inference-time scaling behavior of Base Model, GRPO, and GRPO w/ SDAE.

SDAE improves performance across the entire Pass@k spectrum. As shown in Figure 5, on AIME25, GRPO w/ SDAE consistently outperforms both the base model and vanilla GRPO across all k values (16 to 256). At k=256, SDAE achieves 63.3%, surpassing GRPO (57.4%) and the base model (53.3%). Similar trends are observed on AIME 2024, AMC23, and MATH500. This indicates that SDAE not only improves Pass@1 accuracy but also elevates the model's reasoning ceiling under more attempts.

SDAE preserves exploration capability lost by standard RL. A key finding from Table 2 is that standard RL methods often achieve lower Pass@k at high k values than the base model. On MATH, the base model achieves 96.9% Pass@256, while GRPO drops to 95.5% and PSR drops to 91.2%. In contrast, SDAE variants maintain or exceed the base model's ceiling: REINFORCE w/ SDAE achieves 99.2%, GRPO w/ SDAE achieves 97.2%. On AMC23, both SDAE variants achieve perfect 100.0% Pass@256, while PPO (97.5%) and PSR (92.5%) fail to preserve the base model's full exploration capability.

Semantic diversity achieves better exploration-exploitation balance. The Pass@k curves reveal that SDAE achieves a superior balance: it improves low-k performance (exploitation) without sacrificing high-k performance (exploration). This contrasts with PSR, which aggressively improves Pass@1 but collapses diversity at high k. By weighting advantages based on semantic gravity rather than token-level entropy, SDAE encourages diverse solution paths while still reinforcing correct solutions.

K Entropy Dynamics: Mitigating Strategy Collapse

To further investigate how SDAE prevents models from falling into local optima, we monitor policy entropy throughout training on the DAPO-Math-17K dataset. Policy entropy is a core metric for

measuring model exploration potential. As shown in Figure 6, we compare SDAE against vanilla GRPO and state-of-the-art entropy control methods Clip-Cov and KL-Cov [7].

In vanilla GRPO (light blue curve), we observe sharp and monotonic entropy decline, collapsing to approximately 0.05 within 300 training steps. This "entropy collapse" phenomenon indicates the policy has become overconfident on a very limited set of solution paths, leading to loss of exploration capability and performance saturation.

In contrast, our proposed SDAE (green curve) effectively mitigates this collapse trend, stabilizing average entropy around 0.21. Unlike token-level interventions, SDAE operates at the response level. By considering global semantic structure, SDAE preserves diversity among semantically distinct solutions.

Compared to methods like Clip-Cov (0.22) and KL-Cov (0.20) that explicitly constrain high-covariance token updates, SDAE achieves comparable entropy maintenance without requiring complex per-token covariance computations. This demonstrates that encouraging semantic-level diversity can provide a global exploration signal that is equally effective as specialized gradient-level constraints in maintaining model "vitality" and preventing policy collapse.

L Analysis of Training Dynamics and Behavior Evolution

To deeply investigate SDAE's learning mechanism, we track four metrics during training on the MATH dataset: (1) greedy decoding accuracy on the test set, (2) policy entropy, (3) the ratio of correct responses per batch on the training set, and (4) the proportion of fully-solved prompts per batch (i.e., all rollouts are correct) on the training set. These metrics reveal how different objectives shape the balance between exploration and exploitation, with results shown in Figure 7.

As shown in Figure 7a, SDAE-enhanced models achieve significantly better greedy decoding accuracy than vanilla GRPO and NSR. Although PSR shows faster accuracy growth in early training, it quickly plateaus due to falling into the "pure exploitation" trap (i.e., overfitting to initial correct paths at the expense of generalized reasoning capability). In contrast, SDAE maintains steady upward trends throughout training, demonstrating superior convergence.

Figure 7b shows model exploration potential through policy entropy. Both vanilla GRPO and PSR exhibit severe entropy collapse (dropping below 0.05), indicating loss of generation diversity. While NSR maintains higher entropy by suppressing incorrect samples, its accuracy improvement is slower. SDAE successfully achieves the optimal balance: it maintains significantly higher and stable entropy levels than vanilla GRPO (0.15), preserving sufficient reasoning boundaries while optimizing accuracy, avoiding the policy collapse common in standard RLVR.

The trends in Figure 7c and d indicate that SDAE is more efficient than NSR at discovering correct solution paths. By using semantic gravity to guide exploration, SDAE is more aggressive than pure negative sample reinforcement at improving correct sample ratios and fully-solved ratios. This confirms that rewarding semantic-level diversity is a more potent signal for eliciting latent model reasoning behaviors compared to mere failure penalties.

M Impact of Embedding Spaces on Semantic Exploration

To validate SDAE's robustness and the impact of different embedding spaces on semantic exploration, we compare the default Latent Semantic Aggregation (LSA) against External Semantic Encoders (ESE) using external embedding models of varying scales (Qwen3-Embedding-0.6B and 4B).

As shown in Table 8, the LSA method consistently attains the best or highly competitive performance across both benchmarks. Specifically, on AIME25, LSA achieves 63.3% Pass@256, outperforming the 4B ESE version (60.0%) and 0.6B ESE version (56.7%). This indicates that the policy model's own latent space already adequately models reasoning diversity, and leveraging internal states mitigates potential distribution shifts that may be introduced by misaligned external encoders.

Among ESE variants, we observe clear scaling laws: scaling the external encoder parameters from 0.6B to 4B yields monotonic improvements in Pass@k results. For example, on AMC23, the 4B ESE version achieves 96.0% Pass@64, significantly higher than the 0.6B version's 93.9%.

Table 8: Performance comparison (Pass@k) on AIME25 and AMC23 benchmarks. We compare our SDAE framework using LSA against ESE with varying model scales.The red ✗ indicates that no external model is used. Light blue and light yellow indicate the best performance in each column.

Method	Settings	Pass@k								
		1	2	4	8	16	32	64	128	256
AIME25										
Base Model	✗	6.1	9.7	13.8	17.9	22.2	26.5	30.8	36.6	46.7
GRPO	✗	10.3	14.7	19.4	24.0	28.4	32.8	37.3	42.5	50.0
GRPO w/ SDAE (LSA)	✗	**11.7**	**15.8**	**20.0**	**24.5**	**30.0**	**36.3**	**44.1**	**53.6**	**63.3**
GRPO w/ SDAE (ESE)	Qwen3-Embedding-0.6B	9.6	13.4	18.1	23.2	28.1	33.9	40.9	48.7	56.7
	Qwen3-Embedding-4B	10.6	14.2	19.0	24.0	29.7	**36.3**	43.1	50.4	60.0
AMC23										
Base Model	✗	41.0	56.2	69.2	78.9	85.1	89.1	92.9	97.2	**100.0**
GRPO	✗	61.7	68.7	74.6	80.0	85.1	89.7	93.4	95.9	97.5
GRPO w/ SDAE (LSA)	✗	**62.5**	**70.6**	**77.5**	**83.5**	**88.3**	**92.6**	**96.1**	98.9	**100.0**
GRPO w/ SDAE (ESE)	Qwen3-Embedding-0.6B	60.9	69.7	76.7	81.8	86.1	90.2	93.9	96.5	97.5
	Qwen3-Embedding-4B	61.9	**70.2**	76.6	81.9	86.0	90.8	96.0	**99.3**	**100.0**

Despite different embedding sources, all SDAE configurations significantly outperform the vanilla GRPO baseline at higher k values. LSA's superior performance demonstrates that the "semantic geometry" required for effective exploration is already intrinsically captured in the policy's own reasoning process, making SDAE an efficient and self-contained framework that does not strictly depend on additional heavy external models.

N Analyses of Semantic Diversity and Reasoning Quality

We provide additional evidence that SDAE improves not only final-answer accuracy, but also the semantic diversity and process quality of reasoning. Unless otherwise stated, the experiments in this section use Qwen2.5-Math-7B trained on MATH with either GRPO or GRPO w/ SDAE and evaluated on AIME25.

N.1 Direct Evaluation of Solution Diversity

To directly evaluate solution diversity, we use GPT-5.4-Thinking as an external annotator on all 30 problems in AIME25. For each problem, we sample 8 rollouts and annotate the strategies of the correct rollouts. We report three metrics:

- **Avg. Correct Rollouts:** average number of correct answers among the 8 sampled rollouts.
- **Avg. Distinct Strategy Clusters:** average number of semantically distinct correct solution strategies.
- **Strategy Diversity Ratio (Clusters / Correct):** average ratio between the number of distinct correct strategy clusters and the number of correct rollouts. Higher values indicate that correct solutions are less likely to collapse into a single template.

As shown in Table 9, SDAE increases both the number of correct rollouts and, more importantly, the number of semantically distinct correct strategies. This suggests that SDAE improves successful exploration rather than merely repeating the same correct pattern more frequently.

Table 9: Direct evaluation of strategy-level diversity on AIME25.

Metric	GRPO	GRPO w/ SDAE
AIME25		
Avg. Correct Rollouts	1.2	**1.4**
Avg. Distinct Strategy Clusters	0.43	**0.67**
Strategy Diversity Ratio (Clusters / Correct)	0.36	**0.48**

N.2 Step-Level Reasoning Accuracy

To assess whether SDAE improves the reasoning process itself rather than only the final answer, we ask GPT-5.4-Thinking to judge the logical correctness of each substantive reasoning step in every rollout across the same 30 AIME25 problems.

We report two metrics:

- **Avg. Reasoning Steps:** the average number of substantive reasoning steps per rollout.
- **Step-Level Accuracy (Correct Steps / Total):** the proportion of logically correct steps among all annotated reasoning steps.

As shown in Table 10, the two methods produce chains of thought with comparable lengths, but SDAE attains substantially higher step-level accuracy. This indicates that SDAE improves not only final-answer correctness, but also the quality of intermediate reasoning.

Table 10: Step-level reasoning accuracy on AIME25.

Metric	GRPO	GRPO w/ SDAE
AIME25		
Avg. Reasoning Steps	7.8	**8.0**
Step-Level Accuracy (Correct Steps / Total)	47.9	**55.8**

N.3 Qualitative Case Study on Strategy Diversity

We next present a representative AIME25 example to illustrate how SDAE changes the distribution of successful reasoning strategies.

Problem: *Find the sum of all integer bases* $b > 9$ *for which* 17_b *is a divisor of* 97_b.

As shown in Table 11, GRPO's correct solutions collapse almost entirely to a single dominant strategy, whereas SDAE preserves an additional semantically distinct correct strategy while maintaining the same accuracy. The two dominant correct strategies are:

- **Strategy A (Divisor Enumeration):** Rewrite $17_b = b + 7$,$97_b = 9b + 7$, and observe that $9b + 7 = 9(b + 7) - 56$. Hence, if $17_b \mid 97_b$, then $b + 7 \mid 56$. Together with the constraint $b > 9$, we have $b + 7 > 16$, so $b + 7$ must be a positive divisor of 56 greater than 16, namely 28 or 56. Therefore, $b = 21$ or $b = 49$, and the required sum is $21 + 49 = 70$.
- **Strategy B (Quotient Parameterization):** Set $9b + 7 = k(b + 7)$, which gives $b(9 - k) = 7(k - 1)$. Since $9b + 7 > b + 7$, we must have $k > 1$; moreover, because the right-hand side is positive and $b > 0$, we also need $9 - k > 0$. Hence, $2 \le k \le 8$. Solving for b yields $b = \frac{7(k-1)}{9-k}$. Enumerating integer values $k = 2, 3, \ldots, 8$, we find that only $k = 7 \Rightarrow b = 21, k = 8 \Rightarrow b = 49$ satisfy both integrality and the condition $b > 9$. Therefore, the required sum is again $21 + 49 = 70$.

The key difference between the two strategies is that Strategy A works directly from the divisibility relation and reduces the problem to enumerating divisors of a fixed constant, whereas Strategy B explicitly introduces the quotient parameter k and solves an induced Diophantine constraint over (b, k). The former is organized around factor structure, while the latter is organized around quotient parameterization. Thus, although the two solutions reach the same answer, they correspond to semantically different reasoning paths.

Although both methods solve the problem perfectly, GRPO collapses almost entirely to a single dominant strategy, whereas SDAE consistently preserves an additional semantically distinct correct strategy. This provides direct qualitative evidence that SDAE improves the diversity of successful solution paths.

Table 11: Strategy distribution for one AIME25 example.

Method	Correct rollouts	Strategy A	Strategy B
GRPO	8/8	8	0
GRPO w/ SDAE	8/8	5	3

O Qualitative Example of LSA vs. an External Embedding Model

We also examine a setting in which two responses are mathematically correct but correspond to different reasoning strategies. Consider the problem:

Find the sum of all positive integers n such that $n^2 + 12n - 2007$ is a perfect square.

Two valid solutions are:

- **Solution A (Difference of Squares):** Set $n^2 + 12n - 2007 = m^2$. Complete the square: $(n+6)^2 - m^2 = 2043$. So $(n+6-m)(n+6+m) = 2043$. Factor $2043 = 3^2 \cdot 227$. The positive factor pairs are only $(1, 2043), (3, 681), (9, 227)$. From $n + 6 = \frac{a+b}{2}$, we get $n = \frac{a+b}{2} - 6$, yielding $n = 1016, 336, 112$. The sum is $1016 + 336 + 112 = 1464$.
- **Solution B (Discriminant):** Treat $n^2 + 12n - (2007 + m^2) = 0$ as a quadratic in n. For integer solutions, the discriminant must be a perfect square: $\Delta = 12^2 + 4(2007 + m^2) = 4(m^2 + 2043)$. Set $\Delta = (2t)^2$, then $t^2 - m^2 = 2043$, i.e., $(t-m)(t+m) = 2043$. Factor 2043 into positive factor pairs: $(1, 2043), (3, 681), (9, 227)$. Then $n = \frac{-12+\sqrt{\Delta}}{2} = t - 6$, where $t = \frac{u+v}{2}$, so $n = \frac{u+v}{2} - 6$. This yields $n = 1016, 336, 112$, and the sum is 1464.

As shown in Table 12, LSA assigns a large distance to these two solutions, correctly reflecting the fact that they use different mathematical strategies. By contrast, the external embedding model produces a much smaller distance, likely because it is dominated by shared surface patterns such as similar equations and symbols. This example illustrates why policy-aligned internal representations are more suitable for measuring reasoning diversity than a generic external encoder.

Table 12: Cosine distance between two correct but semantically different solutions.

Representation	**Cosine Distance**
LSA (Ours)	**0.8953**
External embedding model (Qwen3-Embedding-0.6B)	0.1517

P Design Analysis and Controlled Comparisons

Unless otherwise stated, all experiments in this section use Qwen2.5-Math-7B, follow the training setup in Appendices F and G, train on MATH, and evaluate on AIME25.

P.1 Advantage Modulation vs. Reward Shaping and Distance Functions

We compare the following variants:

- **GRPO:** standard GRPO without semantic modulation.
- **Reward shaping + Cosine:** additive reward shaping with the same semantic signal, using $r'_i = (2r_i - 1) \cdot (1 + SG_i)$, followed by standard GRPO advantage computation from the shaped rewards.
- **Advantage Modulation + Euclidean:** replace cosine distance in SDAE with Euclidean distance while keeping the same advantage-modulation framework.
- **Advantage Modulation + Cosine (Ours):** the full SDAE formulation in Eq. 5.

As shown in Table 13, under the same cosine-based semantic signal, advantage modulation clearly outperforms direct reward shaping, with a $+6.6$ improvement at Pass@256. This supports the central design choice of SDAE: the method works best when semantic geometry modulates the contribution of each sample to the policy update, rather than directly altering the reward target.

Table 13: Ablation on modulation form and distance function.

Variant	Pass@1	Pass@16	Pass@64	Pass@256
	AIME25			
GRPO	10.3	28.4	37.3	50.0
+ Reward Shaping	10.9	28.7	40.2	56.7
+ Advantage Modulation + Euclidean	11.4	29.5	43.6	60.0
+ Advantage Modulation + Cosine (Ours)	**11.7**	**30.0**	**44.1**	**63.3**

Under the same advantage-modulation framework, Euclidean distance remains competitive but is consistently weaker than cosine distance. This suggests that SDAE's main benefit comes from the semantic-geometric modulation mechanism itself, while cosine distance offers an additional advantage through scale invariance and robustness to hidden-state norm drift.

P.2 Pairwise Distance Baseline

We further compare the centroid-based SDAE formulation with a direct pairwise-distance baseline. In this baseline, we replace the distance to the centroid of correct responses with the mean pairwise distance among sampled responses, while keeping the same advantage-modulation framework.

As shown in Figure 8, the pairwise formulation is a meaningful baseline, but the centroid-based design is consistently stronger across all Pass@k values. We attribute this to the directional role of the correct-solution centroid: pairwise averaging measures only overall dispersion, whereas the centroid anchors semantic distance to the set of correct solutions and therefore better distinguishes promising novel exploration from irrelevant divergence.

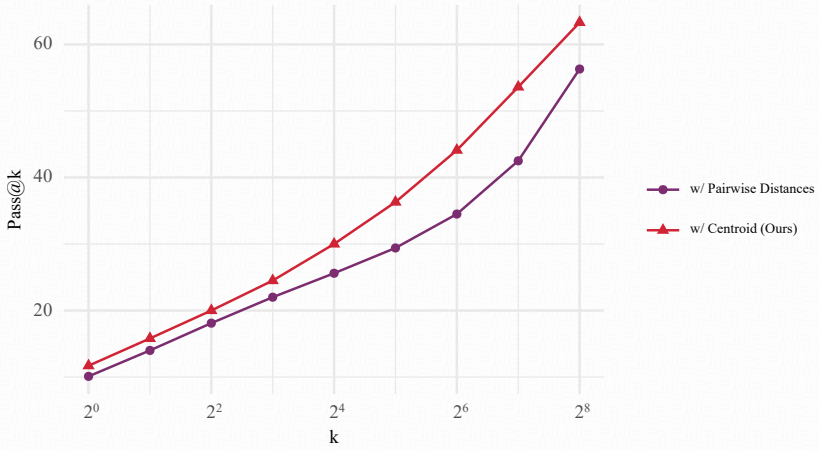


Figure 8: Centroid-based SDAE vs. pairwise-distance baseline on AIME25.

Q Reliability of the Positive Centroid and Effect of Rollout Size

SDAE computes the gravitational source $\boldsymbol{\mu}^+$ from the correct responses within a group of $N = 8$ rollouts. This section discusses what role this in-group centroid plays, how noise in the centroid affects the policy update, what happens when correct responses are sparse, and how the rollout size N is expected to affect SDAE.

Q.1 Role of the Positive Centroid

The centroid $\boldsymbol{\mu}^+$ is not intended to approximate the global mean of all possible correct solutions to a problem; if it were, an estimate from a handful of samples would indeed be insufficient. Instead, it serves as a local, correctness-conditioned anchor that indicates where, in the policy's own representation space, the correct trajectories that the current policy can already produce are located. SDAE then uses the distance of each rollout to this anchor to determine how strongly that rollout contributes to the policy update.

Q.2 Bounded Failure Mode under Centroid Noise

Since the normalized semantic gravity $\hat{SG}_i$ (Appendix C) is non-negative, the modulated advantage satisfies

$$\operatorname{sign}\left(A_i \cdot \hat{SG}_i\right) = \operatorname{sign}(A_i) \quad \text{whenever } \hat{SG}_i > 0, \tag{23}$$

so the original advantage A_i alone determines whether a rollout is reinforced or suppressed, while $\hat{SG}_i$ only rescales the magnitude of its update. Moreover, the mean multiplier within a group is

$$\frac{1}{N}\sum_{i=1}^{N} \hat{SG}_i = \frac{\overline{SG}}{\overline{SG} + \epsilon} \approx 1, \qquad \overline{SG} = \frac{1}{N}\sum_{j=1}^{N} SG_j. \tag{24}$$

Consequently, an inaccurate centroid can perturb the distance ordering and cause some rollouts to be over- or under-weighted, but it cannot turn a verified correct rollout into a negative update or an incorrect rollout into a positive one, and because the mean multiplier stays close to one, such errors mainly redistribute update magnitudes among the rollouts of a group rather than uniformly inflating or shrinking the update of the whole group. The failure mode of a noisy anchor is therefore an imperfect allocation of update magnitudes, not an incorrect optimization direction.

Q.3 Sparse Positive Groups

The most challenging regime is a group in which only a few rollouts are correct, e.g., two correct and six incorrect responses out of $N = 8$. In this case the anchor is inevitably noisier than when more correct responses are available, and its quality depends on the correct responses themselves: if the two correct responses are semantically similar, their centroid provides a relatively coherent anchor; if they follow substantially different strategies, the centroid is less representative of either, and the resulting distance ranking becomes noisier. Even then, the centroid remains conditioned on verified correctness, and by the argument above the resulting error is confined to the magnitudes of the updates. We do not claim, however, that the anchor is equally stable in all cases. Groups with no correct response or with only correct responses are handled separately, as described in Appendix B.

Regarding how often sparse positive groups occur, Figure 7c shows that the per-batch ratio of correct responses is low early in training and increases steadily thereafter, which suggests that sparse positive groups are more common in early training and on difficult prompts. This figure does not, however, directly quantify the frequency of groups with a specific number of correct responses, so we do not treat it as direct evidence for any particular case.

Q.4 Effect of the Rollout Size

All main experiments use $N = 8$ rollouts per prompt. Below we discuss how the rollout size is expected to affect the centroid and the resulting performance.

Centroid stability. Increasing N increases the expected number of correct rollouts available for constructing $\boldsymbol{\mu}^+$ and reduces the frequency of extremely sparse positive groups. Under standard sampling assumptions, averaging more correct representations reduces the variance of the centroid estimate. A larger budget such as $N = 16$ would therefore likely provide a more stable anchor than $N = 8$, particularly for difficult prompts with few correct rollouts.

Expected effect on performance. A larger N also strengthens the baseline itself, since GRPO obtains more samples for group-wise advantage normalization; the relevant question is therefore whether SDAE's improvement over the baseline grows with N. We expect this improvement to be non-negative, because the positive/negative semantic structure that SDAE exploits is characterized

more fully with more samples. At the same time, SDAE depends on the relative distances of rollouts to the anchor rather than on the anchor's absolute position, and we expect these relative distances to be less sensitive to moderate perturbations of the centroid, which is consistent with the substantial gains already obtained at $N = 8$. Taken together, we expect a larger rollout budget to yield a consistent but bounded additional improvement, and we leave a systematic study to future work.

Why $N = 8$. This choice is driven primarily by comparability. The methods we compare against, including GRPO and Entropy Adv [10] as well as PSR, NSR, and W-REINFORCE [12], all use 8 rollouts per prompt. Using a larger budget for SDAE alone would confound the effect of the method with a higher sampling cost. Obtaining improvements under an identical rollout budget indicates that SDAE's benefit comes from making better use of the existing samples rather than from sampling more.

R Choice of Layer for Latent Semantic Aggregation

LSA pools the last-layer hidden states of the policy model (Eq. 1). To examine whether this choice matters, we compare it with representations taken from intermediate layers of the same policy. Following the setup of Appendix P, we train Qwen2.5-Math-7B with GRPO w/ SDAE on MATH and evaluate on AIME25, keeping all other settings fixed and changing only the layer from which $\mathbf{h}_{i,t}$ in Eq. 1 is taken:

- **Intermediate layer:** hidden states of a single representative intermediate layer.
- **Intermediate average:** hidden states averaged across multiple intermediate layers.
- **Last layer (Ours):** the default LSA.

Table 14: Pass@k on AIME25 with hidden states pooled from different layers of the policy model.

Pooling source	Pass@1	Pass@2	Pass@4	Pass@8	Pass@16	Pass@32	Pass@64	Pass@128	Pass@256
				AIME25					
GRPO	10.3	14.7	19.4	24.0	28.4	32.8	37.3	42.5	50.0
+ SDAE, Intermediate layer	11.0	15.4	18.9	23.1	26.3	31.9	39.6	45.1	50.0
+ SDAE, Intermediate average	11.4	15.0	19.1	23.0	27.2	32.2	40.1	47.0	53.3
+ SDAE, Last layer (Ours)	**11.7**	**15.8**	**20.0**	**24.5**	**30.0**	**36.3**	**44.1**	**53.6**	**63.3**

As shown in Table 14, last-layer pooling performs best at every k, and the gap generally widens as k grows: at Pass@1 the three variants differ by at most 0.7 points, whereas at Pass@256 last-layer pooling reaches 63.3% compared with 50.0% and 53.3% for the two intermediate variants. With intermediate-layer representations, SDAE's gains over GRPO largely disappear: both intermediate variants fall slightly below GRPO for $4 \leq k \leq 32$, are at most 4.5 points above it for $k \geq 64$, and the single intermediate layer only matches GRPO at Pass@256. They also fall below both external-encoder variants in Table 8 at Pass@256 (56.7% and 60.0%), suggesting that the benefit of LSA comes not from using internal representations per se, but specifically from the last layer.

This is consistent with the analysis in Appendix A: the last layer is the bottleneck through which all trajectory-specific first-order optimization information must pass, so distances measured there are most closely aligned with the geometry of the policy update. We hypothesize that intermediate layers carry more generic, optimization-agnostic features, which yield a noisier modulation signal. The result also delimits the scope of SDAE: its benefit presupposes a semantic space aligned with the optimization geometry, and replacing that space with a less aligned one largely removes the gain.

S Limitations

While SDAE demonstrates consistent improvements across multiple benchmarks, several limitations warrant discussion and point to promising directions for future research.

Dependence on Correct Samples for Gravitational Source. SDAE relies on the centroid of correct responses to serve as the gravitational source. For extremely challenging problems where few or no correct rollouts are sampled, the method must fall back to using the consensus of all responses (as discussed in Appendix B). While this fallback preserves some discriminative capacity between local and divergent errors, the exploration guidance may be less effective compared to scenarios with

well-defined positive centroids. Investigating adaptive strategies for gravitational source estimation under sparse correctness signals remains an open problem.

Domain Generalization. Our experiments focus exclusively on mathematical reasoning tasks, which benefit from deterministic verification and well-structured solution spaces. The effectiveness of SDAE on other reasoning domains, such as code generation, commonsense reasoning, or open-ended tasks where correctness is less binary, remains to be validated. In particular, tasks with subjective or multi-faceted evaluation criteria may require modified formulations of semantic gravity.

Sensitivity to Internal Representation Quality. SDAE relies on Latent Semantic Aggregation (LSA) to extract geometric information from the model's last-layer hidden states. If the base model's pre-trained embedding space is insufficiently developed or exhibits high noise in specific domains, the resulting semantic gravity may not accurately reflect the true reasoning diversity, potentially leading to biased advantage modulation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the motivation, the proposed SDAE method, and the scope of the empirical claims. These claims are supported by the method section and the experimental results in Experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly discusses limitations in Appendix S.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We provide theoretical results in Section 3.3 and Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper discloses the datasets, backbones, baselines, evaluation benchmarks, and decoding settings in Section 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All the datasets used in our paper are open-access, and we also provide our codes in the supplemental materials.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the core experimental setting in Section 4.1, including datasets, models, baselines, metrics, and decoding configurations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Our results are based on 256 random samplings and calculated with an unbiased estimator.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the training framework and optimization setup in Section 4.1 and Appendix G, and it further analyzes the computational and memory overhead of SDAE in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research was conducted in accordance with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Our work is a methodological and theoretical study of reinforcement learning for improving large language model reasoning, and it does not have direct societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk generative model, scraped dataset, or other asset requiring controlled access or special safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original sources of the datasets, models, and prior methods used in this work. Their corresponding licenses and terms of use are respected in our implementation and release materials.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release our code, used datasets and provided detailed documentation in the supplementary materials.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not an important, original, or non-standard component of the core SDAE method itself. Although GPT-5.4-Thinking is used as an auxiliary annotator in an appendix analysis, it is not part of the proposed training objective or the main experimental pipeline.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.