Free Energy-Driven Reinforcement Learning with Adaptive Advantage Shaping for Unsupervised Reasoning in LLMs

Anonymous ACL submission

Abstract

Unsupervised reinforcement learning (RL) has emerged as a promising paradigm for enabling self-improvement in large language models (LLMs). However, existing unsupervised RL-based methods often lack the capacity to adapt to the model's evolving reasoning capabilities during training. Therefore, these methods can misdirect policy optimization in the absence of ground-truth supervision. To address this issue, we introduce **FREIA**, a novel RL-based algorithm built on two key innovations: (1) *Free Energy-Driven Reward (FER)* adapts rewards to balance consensus and exploration based on the Free Energy Principle. (2) *Adaptive Advantage Shaping (AAS)* adaptively adjusts learning signals based on the statistical characteristics of sampled rewards. Empirical evaluations on nine datasets across three reasoning tasks showcase that FREIA outperforms other unsupervised RL-based baselines. Notably, in mathematical reasoning tasks, FREIA surpasses other methods by an average of 0.5 to 3.5 points in Pass@1 using the DeepSeek-R1-Distill-Qwen-1.5B model.

1 Introduction

Reinforcement Learning with Verifiable Rewards (RLVR) (Wen et al., 2025) has emerged as a fundamental technique for enhancing the reasoning capabilities of Large Language Models (LLMs) (Bai et al., 2025; Guo et al., 2025). Through alignment with ground-truth supervision signals, RLVR has demonstrated substantial improvements across a wide range of reasoning tasks (Chen et al., 2025; Wang et al., 2025; Yu et al., 2018). Nevertheless, this paradigm critically depends on external supervision, which imposes significant data annotation costs requiring human expertise. As a result, recent research has increasingly explored methods that enable models to learn from intrinsic signals, a process referred to as *unsupervised self-improvement* (Zuo et al., 2025).

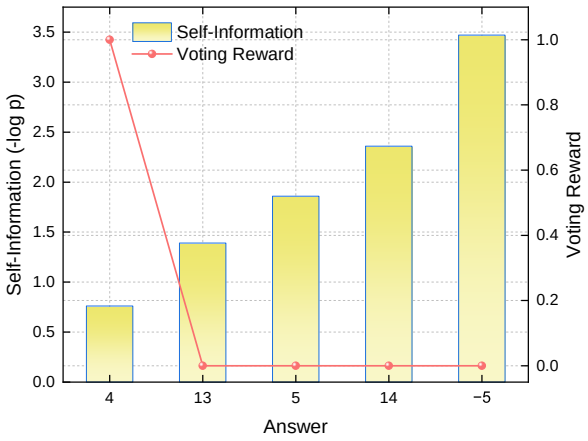


Figure 1: An analysis of reward signals for a math problem, where the correct answer is "13". Note that p indicates the frequency of sampled answers.

A key challenge in promoting unsupervised self-improvement lies in the formulation of reliable learning signals. To cope with this, existing unsupervised methods can be categorized into *trajectory-intrinsic methods* and *population-based methods*. Specifically, *trajectory-intrinsic methods* derive rewards from intrinsic measures of reasoning paths such as model uncertainty (Prabhudesai et al., 2025; Li et al., 2025; Agarwal et al., 2025; Zhao et al., 2025; Zhang et al., 2025a). In contrast, *population-based methods* produce reward signals by comparing each reasoning output against a set of candidate outputs to exploit collective consensus. They assume that majority agreement reflects correctness (Prasad et al., 2024; Zuo et al., 2025; Liu et al., 2025; Yuan et al., 2025; Zhang et al., 2025b). However, both paradigms cannot adapt to the model's evolving capabilities during training, thereby undermining effective policy optimization.

First, current methods often misdirect policy optimization by applying static criteria. Specifically, *population-based methods* rely solely on consensus. They discard valuable information from minority reasoning paths and impede necessary ex-

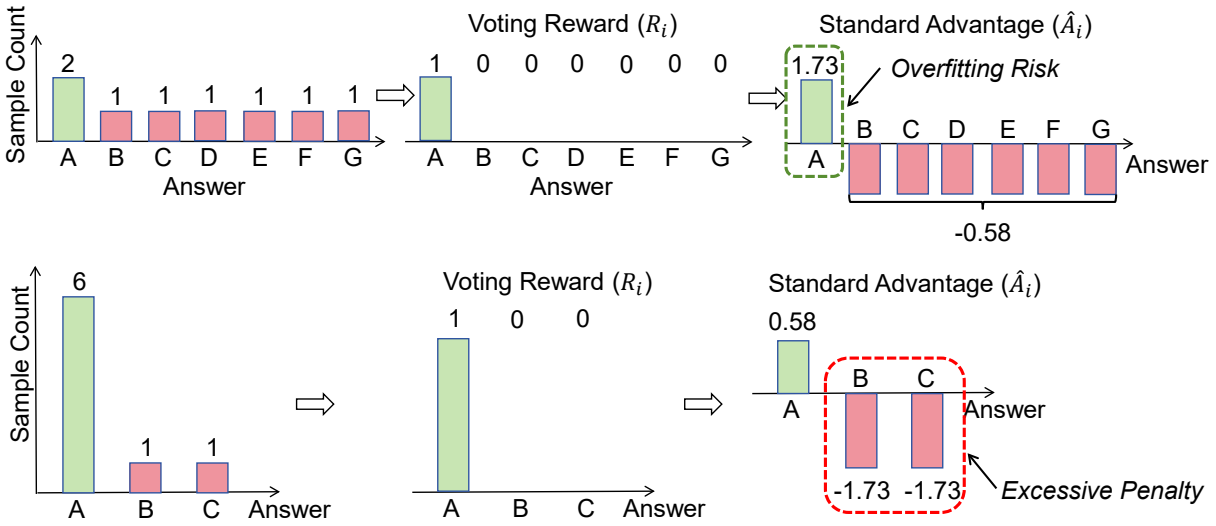


Figure 2: An analysis of standard advantage shaping. **(Top: Weak Consensus)** The high-reward answer appears infrequently. Standard advantage assigns excessive values to these rare answers. **(Bottom: Strong Consensus)** The majority answer dominates the population. Standard advantage excessively penalizes the occasional deviations.

ploration, particularly during early training stages. As shown in Figure 1, majority voting assigns zero reward to the correct answer ("13") while granting the maximum reward to the incorrect majority ("4"). This bias reinforces erroneous consensus and increases vulnerability to incorrect agreement in unsupervised settings (Huang et al., 2025). In contrast, *trajectory-intrinsic methods* risk reinforcing high-confidence but incorrect answers. This risk is particularly pronounced when the model's reasoning ability is limited, leading to a significant mismatch between model confidence and actual correctness (Zhang et al., 2025a). Figure 1 illustrates that approaches relying solely on self-information assign high rewards to rare answers without considering correctness. This bias disrupts robust policy optimization by rewarding incorrect solutions[1].

Beyond these design flaws, a second critical issue emerges in how these rewards are processed. Existing unsupervised RL-based methods apply static advantage shaping despite the continuous changes of reward distributions during training. As depicted in Figure 2 (Top), training often begins in a *Weak Consensus* phase, where high-reward answers are sparse. In this phase, standard advantage normalization assigns large positive advantages to infrequent high-reward samples. Without label verification, this bias risks overfitting to potentially misleading signals and early convergence before exploring all possible solutions. As the model reaches a *Strong Consensus* (Figure 2, Bottom), the same strategy assigns excessive penalties to occasional low-reward samples. This shifts the policy optimization focus from refining the dominant consensus to merely avoiding errors.

As a result, addressing these issues necessitates solving a key research problem:

Key Research Problem

How to design an unsupervised RL-based algorithm that adaptively regulates policy optimization to align with the model's evolving learning dynamics?

To answer this question, we propose a new algorithm: *Free Energy-Driven Reinforcement Learning with Adaptive Advantage Shaping* (**FREIA**). This algorithm is designed to enhance the reasoning capabilities of LLMs without external supervision. FREIA integrates two key innovations: (1) **Free Energy-Driven Reward (FER)** formulates a unified reward based on the *Free Energy Principle (FEP)*. FER views self-improvement as a process of minimizing free energy. The objective is decomposed into consensus alignment and exploration of novel reasoning paths. By optimizing this objective, the model leverages its internal uncertainty to mitigate early convergence while reinforcing high-quality reasoning paths as confidence grows. (2) **Adaptive Advantage Shaping (AAS)** adjusts policy updates based on the model's evolving learning dynamics. By analyzing the skewness of re-

[1]The problem is to find the integer $0 \leq n < 18$ such that $n \equiv -11213141 \pmod{18}$.

ward distribution, AAS identifies the current learning phase and adaptively adjusts advantage signals. This approach reduces the impact of unreliable outliers during early exploration and prevents strict penalties for occasional reasoning paths during convergence. Collectively, these innovations enable FREIA to achieve effective self-improvement in unsupervised settings. We evaluate FREIA on nine benchmarks covering three reasoning tasks. Extensive experiments demonstrate its superiority over other unsupervised baselines.

In summary, this work delivers the following contributions:

(1) We identify a fundamental misalignment in existing unsupervised RL-based methods. First, current reward designs fail to adapt to the model's evolving reasoning capabilities. Second, current advantage estimation ignores the shifting distribution of reward signals during training.

(2) We introduce FREIA, a novel unsupervised RL-based algorithm featuring two key innovations. *Free Energy-Driven Reward (FER)* balances consensus and exploration based on the Free Energy Principle. Moreover, *Adaptive Advantage Shaping (AAS)* adaptively modulates advantage estimation based on real-time reward distributions.

(3) Extensive experimental results confirm that FREIA surpasses other unsupervised baselines on nine benchmarks across three reasoning tasks. This showcases its efficacy in enhancing reasoning performance without external supervision.

2 Related Work

Recent efforts to enable self-improvement in LLMs without ground-truth supervision have catalyzed research into unsupervised RL. Existing methods fall into two main paradigms: *trajectory-intrinsic methods* and *population-based methods*.

Trajectory-Intrinsic Methods. These methods assign rewards to individual reasoning paths using intrinsic metrics (*e.g.*, semantic entropy) to estimate self-confidence (Agarwal et al., 2025; Li et al., 2025; Prabhudesai et al., 2025; Zhao et al., 2025; Zhang et al., 2025a). While these approaches aim to reduce uncertainty, they often suffer from self-reinforcement bias. Notably, high-confidence but incorrect answers often receive large rewards in the absence of external verification or peer comparison. Additionally, when model confidence is miscalibrated, these methods may exacerbate systematic errors by reinforcing flawed reasoning patterns. They also lack mechanisms to distinguish between genuine certainty derived from accurate reasoning and overconfidence caused by erroneous correlations, leading to instability throughout the training process.

Population-Based Methods. These methods evaluate answers by comparing each output against other sampled candidates, under the assumption that collective agreement correlates with correctness. Existing strategies can be divided into two levels of granularity: *Final-Answer Consensus* relies on majority voting over final outputs (Prasad et al., 2024; Zuo et al., 2025; Liu et al., 2025; Yuan et al., 2025). This approach often produces stable but rigid learning signals, ignoring valuable minority paths that can lead to correct solutions in later training stages. In contrast, *Cross-View Consensus* (Zhang et al., 2025b) requires agreement across perturbed input variants to improve robustness. However, it often suppresses exploratory behaviors that are critical in complex reasoning tasks. More importantly, both approaches remain outcome-centric, assigning rewards only based on final agreement.

This analysis underscores the need for a unified and principled framework that integrates the strengths of both paradigms while mitigating their respective limitations. Accordingly, this enables more reliable and adaptive model learning in the absence of ground-truth supervision.

3 Methodology

As shown in Figure 3 and Appendix B, FREIA integrates Free Energy-Driven Reward (FER) and Adaptive Advantage Shaping (AAS).

3.1 Free Energy-Driven Reward (FER)

Existing unsupervised RL-based methods typically produce single-dimensional rewards that misalign with the model's learning dynamics. These methods lack the flexibility to adjust the balance between consensus and exploration as the model improves. To address this, we propose **Free Energy-Driven Reward (FER)**, which is grounded in the *Free Energy Principle (FEP)* (Friston, 2010; Buckley et al., 2017). Specifically, FER formulates self-improvement as minimizing free energy. This objective is decomposed into consensus alignment and exploration of novel reasoning paths.

Specifically, given an input x, the model samples a set of G reasoning paths $Y = \{y_1, y_2, \ldots, y_G\}$ and extracts final answers $A = \{a_1, ..., a_G\}$. Let

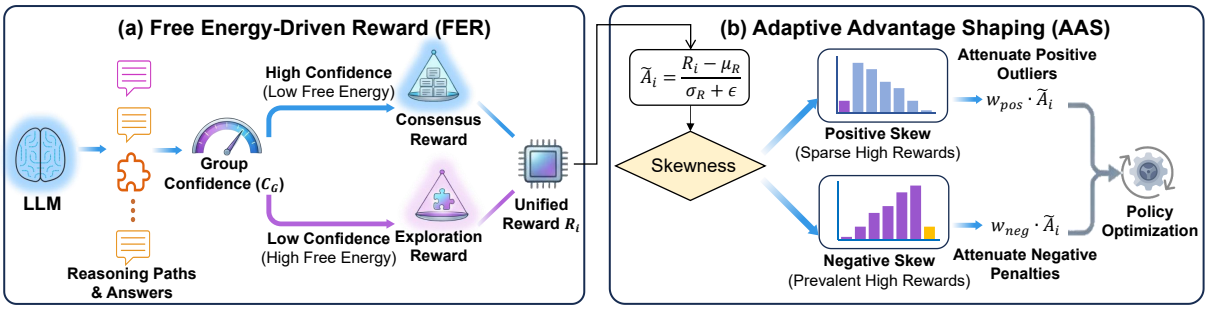


Figure 3: The overall framework of FREIA, including Free Energy-Driven Reward (FER) and Adaptive Advantage Shaping (AAS).

$U = \{u_1, u_2, \ldots, u_M\}$ denote the set of unique answers within A. The frequency distribution of these unique answers $D = \{f_1, f_2, \ldots, f_M\}$ serves as empirical input for the subsequent steps.

Step 1: Each unique answer u_i is regarded as a hypothesis about the ground truth. To distinguish robust consensus from stochastic noise, we apply a *Non-linear Belief Sharpening* mechanism. Using the frequency f_i, the belief weight w_i is given as:

$$w_i = \text{Softmax}(\alpha \cdot \log(f_i)) = \frac{f_i^\alpha}{\sum_{k=1}^{M} f_k^\alpha} \quad (1)$$

where $W = \{w_1, ..., w_M\}$ constitutes the refined belief distribution over candidate answers.

Analysis. Since the true ground-truth distribution is inaccessible in unsupervised settings, an *empirical belief* is derived from the sampled response set. From the perspective of FEP, α modulates the model's confidence in the current batch, forcing the distribution to focus on the leading consensus when α increases. Further analysis on α is shown in Appendix C.4.

Step 2: We define *Group Confidence* C_G to quantify the degree of group consensus for each training sample:

$$C_G = \begin{cases} 1.0 & \text{if } M = 1 \\ 1 - \frac{H(W)}{\log M} = 1 - \frac{-\sum_j w_j \log w_j}{\log M} & \text{if } M > 1 \end{cases} \quad (2)$$

where $H(\cdot)$ denotes the Shannon entropy. $C_G \to 1$ signifies high certainty, whereas $C_G \to 0$ indicates high uncertainty. The superiority of C_G is further examined in Appendix C.2.

Step 3: The total reward for each y_i is formulated as an adaptive trade-off between *Consensus* and *Exploration*, modulated by the group confidence C_G. Specifically, *Consensus* encourages the model to align with the most probable solution:

$$r_{\text{cons}}(y_i) = \begin{cases} 1.0 & \text{if } a_i = \text{Vote}(\{a_k\}_{k=1}^{G}) \\ 0.0 & \text{otherwise} \end{cases} \quad (3)$$

Moreover, *Exploration* incentivizes the discovery of diverse solutions when uncertainty is high. For a given y_i producing answer a_i with belief weight w_i, the reward is defined as:

$$r_{\text{explore}}(y_i) = \tanh(-\log w_i) \quad (4)$$

where the $\tanh$ function serves as a soft normalization mechanism. This strictly bounds r_{explore} within $(0, 1)$, preventing the exploration signal from dominating the policy optimization process.

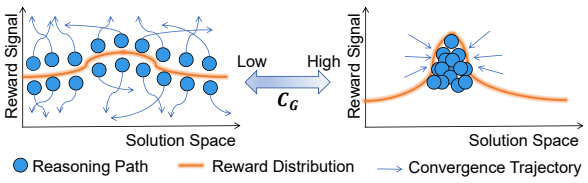


Figure 4: Visualization of FER. Specifically, the left part depicts the Exploration term when C_G is low, while the right part shows the Consensus term under high C_G.

As illustrated in Figure 4, the final reward integrates the *Consensus* and *Exploration* components modulated by C_G, which is given as:

$$R_i = C_G \cdot r_{\text{cons}}(y_i) + (1 - C_G) \cdot r_{\text{explore}}(y_i) \quad (5)$$

Eq. (5) adaptively regulates the learning process, which prioritizes consolidating the consensus in high-confidence states ($C_G \to 1$), while encouraging exploration in low-confidence states ($C_G \to 0$). Further analysis of FER is shown in Appendix C.1.

3.2 Adaptive Advantage Shaping (AAS)

Standard advantage functions are suboptimal for self-improvement in unsupervised settings. Therefore, we propose **Adaptive Advantage Shaping**

(AAS), which adaptively modulates advantages based on the reward distributional characteristics.

Step 1: For each y_i, the standard advantage $\tilde{A}_i$ is derived from the FER scores $\{R(y_k)\}_{k=1}^{G}$:

$$\tilde{A}_i = \frac{R_i - \mu_R}{\sigma_R + \epsilon} \tag{6}$$

where μ_R and σ_R are the mean and standard deviation of the FER scores, respectively. ϵ is a small constant for numerical stability.

Step 2: The learning dynamics of the model are quantified by the *skewness* of the reward distribution $\{R(y_k)\}_{k=1}^{G}$. A positive skew indicates a distribution dominated by low rewards, whereas a negative skew signifies a prevalence of high rewards. The sample skewness $\mathcal{S}$ is defined as (See Appendix C.5 for additional analysis):

$$\mathcal{S} = \frac{1}{G}\sum_{i=1}^{G}\left(\frac{R_i - \mu_R}{\sigma_R + \epsilon}\right)^3 \tag{7}$$

Step 3: Adaptive weights for positive and negative advantages are computed from $\mathcal{S}$ to mitigate potential biases in standard normalization:

$$w_{\text{pos}} = \sigma(-\mathcal{S}),\ w_{\text{neg}} = \sigma(\mathcal{S}) \tag{8}$$

where $\sigma(\cdot)$ denotes the sigmoid function.

Analysis. AAS functions as a stabilizer against unsupervised distributional noise:

Case 1: Positive Skew. In this case, standard normalization assigns excessively high advantages to rare winning paths. However, rewards are computed without ground-truth labels in unsupervised settings, indicating that high rewards do not necessarily correspond to correct or generalizable reasoning. Since such paths may represent stochastic outliers rather than robust solutions, AAS *attenuates positive advantages* ($w_{\text{pos}} \to 0$). This ensures a cautious policy update, preventing the model from overfitting to potentially misleading signals.

Case 2: Negative Skew. In this case, standard normalization assigns large penalties to occasional deviations. Under unsupervised conditions, the absence of ground-truth supervision also makes it unclear whether those deviations are genuine reasoning errors or harmless variations. As these deviations may not reflect fundamental mistakes, AAS *attenuates negative advantages* ($w_{\text{neg}} \to 0$). This ensures that the model does not over-correct based on rare low-reward samples. Additional analysis on AAS is shown in Appendix C.3.

Step 4: The final advantage $\hat{A}_i$ used in the policy update is modulated by these adaptive weights:

$$\hat{A}_i = \begin{cases} w_{\text{pos}} \cdot \tilde{A}_i & \text{if } \tilde{A}_i > 0 \\ w_{\text{neg}} \cdot \tilde{A}_i & \text{if } \tilde{A}_i < 0 \end{cases} \tag{9}$$

This shaped advantage $\hat{A}_i$ is integrated into the standard GRPO objective (Shao et al., 2024):

$$\begin{aligned}\mathcal{L}(\theta) =& \mathbb{E}\Big[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)\hat{A}_i, \\ & \text{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_i\Big) \\ & - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})\Big]\end{aligned} \tag{10}$$

4 Experiments

Our evaluation focuses on four research questions: **RQ1:** Can FREIA achieve superior reasoning performance compared to other unsupervised RL-based baselines? **RQ2:** Does FREIA effectively balance the trade-off between exploration and consensus? **RQ3:** What are the specific contributions of FER and AAS to the overall effectiveness of FREIA? **RQ4:** How sensitive is FREIA's performance to its key hyperparameter α?

4.1 Settings

Models. To assess mathematical reasoning capabilities, we selected Qwen2.5-Math-1.5B-Instruct (Yang et al., 2024b), Qwen2.5-3B-Instruct (Yang et al., 2024a), and DeepSeek-R1-Distill-Qwen-1.5B (Guo et al., 2025). Moreover, Qwen2.5-Coder-3B-Instruct (Hui et al., 2024) and Qwen2.5-VL-3B-Instruct (Bai et al., 2025) were used in SQL generation and multi-modal reasoning, respectively.

Datasets. Training for mathematical reasoning was conducted using MATH (Hendrycks et al., 2021), followed by evaluations on MATH500 (Hendrycks et al., 2021), AIME24 (Li et al., 2024a), AIME25 (Codeforces), AMC23 (Ouyang et al., 2022), Minerva (Lewkowycz et al., 2022), and OlympiadBench (Huang et al., 2024). For SQL generation, we utilized BIRD-Train (Li et al., 2024b) for training, with evaluation performed on Spider-Dev (Yu et al., 2018) and BIRD-Dev (Li et al., 2024b). We used the training and testing splits of Geometry3K (Lu et al., 2021) for multi-modal reasoning.

Implementation Details. In this work, the coefficient of KL loss term was $\beta = 0.001$. The batch size and the number of rollouts were 512 and 8, and

Dataset	Base	GRPO (Supervised)	TTRL	Entropy	Intuitor	FREIA
Qwen2.5-Math-1.5B-Instruct						
MATH500	74.2	$75.2_{\pm 0.7}$	$75.0_{\pm 0.4}$	$74.6_{\pm 0.8}$	$74.4_{\pm 0.8}$	$\mathbf{75.4}_{\pm 0.2}$
AIME24	10.0	$\mathbf{13.3}_{\pm 1.9}$	$\mathbf{13.3}_{\pm 0.0}$	$10.0_{\pm 0.0}$	$10.0_{\pm 0.0}$	$\mathbf{13.3}_{\pm 0.0}$
AIME25	3.3	$\mathbf{16.7}_{\pm 0.0}$	$\mathbf{16.7}_{\pm 0.0}$	$13.3_{\pm 0.0}$	$10.0_{\pm 0.0}$	$\mathbf{16.7}_{\pm 0.0}$
AMC23	47.5	$\mathbf{52.5}_{\pm 1.4}$	$\mathbf{52.5}_{\pm 0.0}$	$47.5_{\pm 0.0}$	$47.5_{\pm 0.0}$	$\mathbf{52.5}_{\pm 0.0}$
Minerva	28.7	$31.3_{\pm 0.9}$	$30.9_{\pm 0.5}$	$29.4_{\pm 0.8}$	$28.7_{\pm 0.9}$	$\mathbf{32.0}_{\pm 0.4}$
Olympiad	35.2	$40.8_{\pm 0.6}$	$40.1_{\pm 0.4}$	$39.2_{\pm 0.7}$	$38.4_{\pm 0.7}$	$\mathbf{41.2}_{\pm 0.3}$
Avg.	33.2	$38.3_{\pm 0.9}$	$38.1_{\pm 0.2}$	$35.7_{\pm 0.4}$	$34.8_{\pm 0.4}$	$\mathbf{38.5}_{\pm 0.2}$
Qwen2.5-3B-Instruct						
MATH500	62.0	$66.0_{\pm 0.6}$	$\mathbf{66.6}_{\pm 0.5}$	$64.6_{\pm 0.7}$	$64.0_{\pm 0.7}$	$65.2_{\pm 0.3}$
AIME24	0.0	$\mathbf{10.0}_{\pm 0.0}$	$6.7_{\pm 0.0}$	$3.3_{\pm 1.9}$	$3.3_{\pm 0.0}$	$\mathbf{10.0}_{\pm 0.0}$
AIME25	0.0	$\mathbf{10.0}_{\pm 0.0}$	$6.7_{\pm 0.0}$	$3.3_{\pm 0.0}$	$3.3_{\pm 0.0}$	$\mathbf{10.0}_{\pm 0.0}$
AMC23	35.0	$37.5_{\pm 1.4}$	$\mathbf{40.0}_{\pm 0.0}$	$37.5_{\pm 0.0}$	$37.5_{\pm 0.0}$	$37.5_{\pm 0.0}$
Minerva	24.3	$25.4_{\pm 0.8}$	$25.8_{\pm 0.5}$	$25.0_{\pm 0.9}$	$24.3_{\pm 0.9}$	$\mathbf{26.1}_{\pm 0.4}$
Olympiad	29.1	$31.5_{\pm 0.6}$	$31.2_{\pm 0.4}$	$30.0_{\pm 0.7}$	$29.8_{\pm 0.8}$	$\mathbf{31.9}_{\pm 0.3}$
Avg.	25.1	$\mathbf{30.1}_{\pm 0.6}$	$29.5_{\pm 0.2}$	$27.3_{\pm 0.7}$	$27.0_{\pm 0.4}$	$\mathbf{30.1}_{\pm 0.2}$
DeepSeek-R1-Distill-Qwen-1.5B						
MATH500	77.6	$82.4_{\pm 0.4}$	$\mathbf{82.6}_{\pm 0.3}$	$81.8_{\pm 0.6}$	$81.4_{\pm 0.6}$	$82.2_{\pm 0.2}$
AIME24	16.7	$\mathbf{20.0}_{\pm 0.0}$	$\mathbf{20.0}_{\pm 0.0}$	$16.7_{\pm 0.0}$	$16.7_{\pm 0.0}$	$\mathbf{20.0}_{\pm 0.0}$
AIME25	16.7	$\mathbf{20.0}_{\pm 1.9}$	$\mathbf{20.0}_{\pm 0.0}$	$16.7_{\pm 0.0}$	$16.7_{\pm 0.0}$	$\mathbf{20.0}_{\pm 0.0}$
AMC23	62.5	$70.0_{\pm 0.0}$	$70.0_{\pm 0.0}$	$65.0_{\pm 0.0}$	$65.0_{\pm 0.0}$	$\mathbf{72.5}_{\pm 0.0}$
Minerva	27.6	$30.5_{\pm 0.6}$	$30.9_{\pm 0.4}$	$29.8_{\pm 0.7}$	$29.4_{\pm 0.7}$	$\mathbf{31.3}_{\pm 0.3}$
Olympiad	42.4	$48.6_{\pm 0.5}$	$49.0_{\pm 0.3}$	$47.5_{\pm 0.6}$	$46.6_{\pm 0.6}$	$\mathbf{49.4}_{\pm 0.3}$
Avg.	40.6	$45.3_{\pm 0.6}$	$45.4_{\pm 0.2}$	$42.7_{\pm 0.3}$	$42.4_{\pm 0.3}$	$\mathbf{45.9}_{\pm 0.1}$

Table 1: Experimental results on multiple mathematical reasoning benchmarks. The results are reported as mean and standard deviation across 3 random seeds (*i.e.*, $\text{Mean}_{\pm \text{Std}}$). The best results are highlighted in bold.

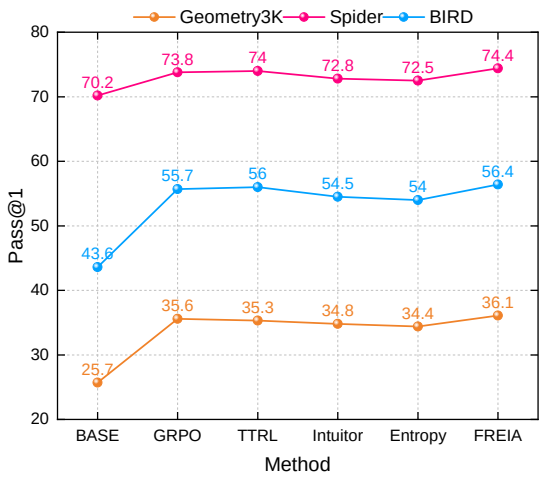


Figure 5: Experimental results on SQL generation and multi-modal reasoning.

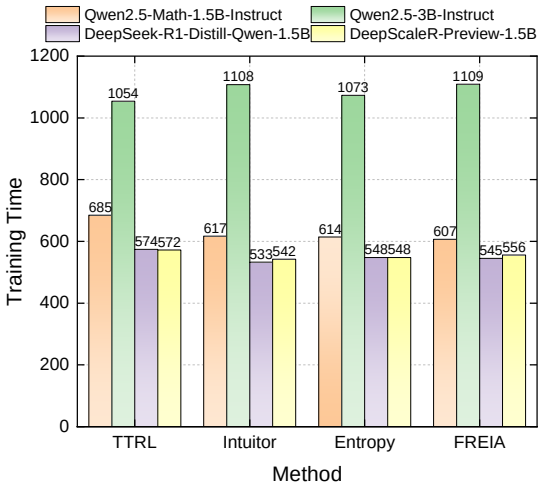


Figure 6: Comparison of total training wall-clock time (minutes) using various methods.

we used a sampling temperature of 1.0. We adopted AdamW (Zhou et al., 2024) with a learning rate of 1×10^{-6} and trained for 400 steps. For FREIA, the parameter α was 2. During evaluation, the sampling temperature was 0.6, and Pass@1 was used for evaluation. We performed experiments using four NVIDIA GeForce A100 40GB GPUs. FREIA was compared against base model, GRPO (supervised RL) (Shao et al., 2024), TTRL (Zuo et al., 2025), Entropy (Prabhudesai et al., 2025), and Intuitor (Zhao et al., 2025), which were set to the same hyperparameters as FREIA. Further experimental setups are shown in Appendix A.

4.2 Experimental Results

4.2.1 Main Results (RQ1)

As shown in Table 1, FREIA achieves superior average performance on mathematical reasoning.

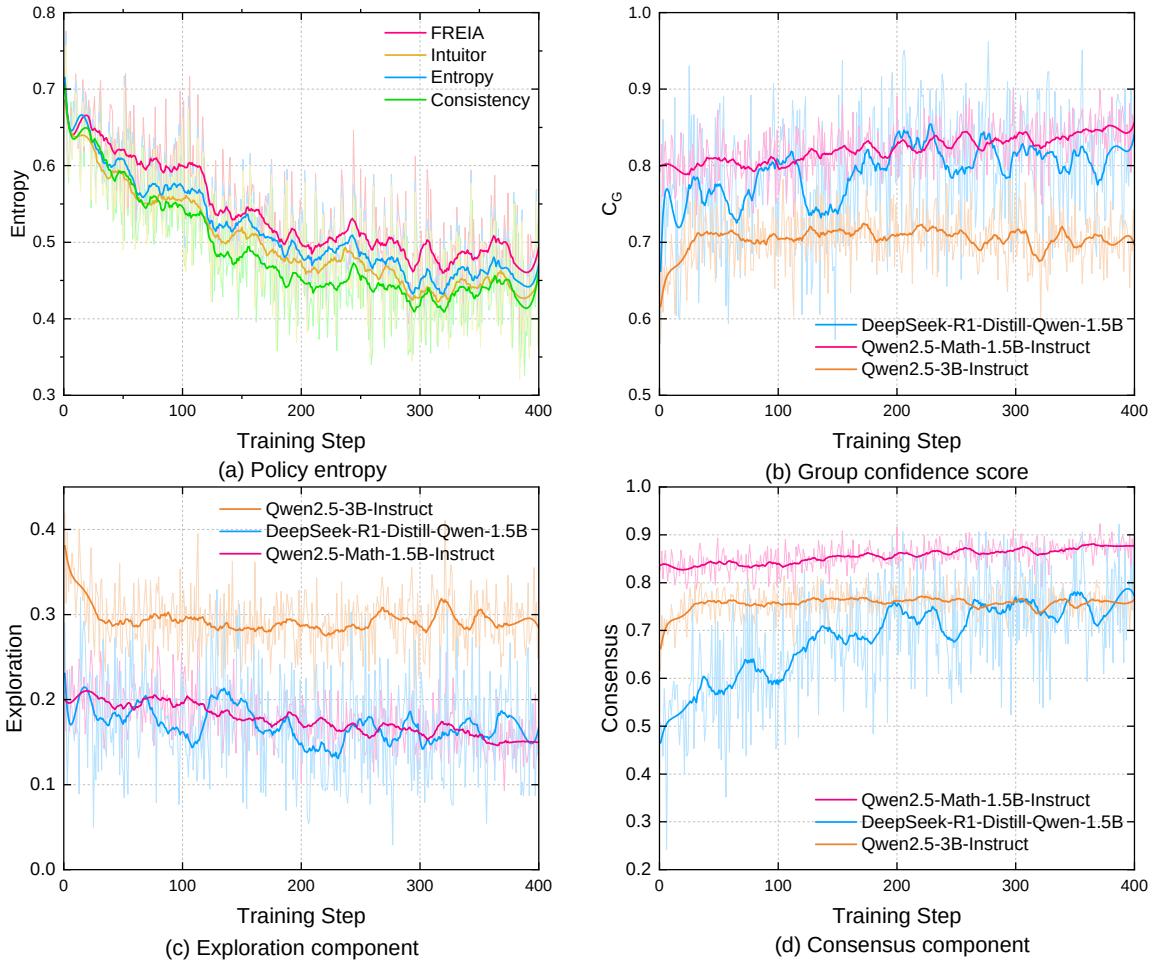


Figure 7: Training dynamics of FREIA. (a) Policy entropy using DeepSeek-R1-Distill-Qwen-1.5B; (b) Group confidence score (C_G); (c) Exploration component of the reward; (d) Consensus component of the reward.

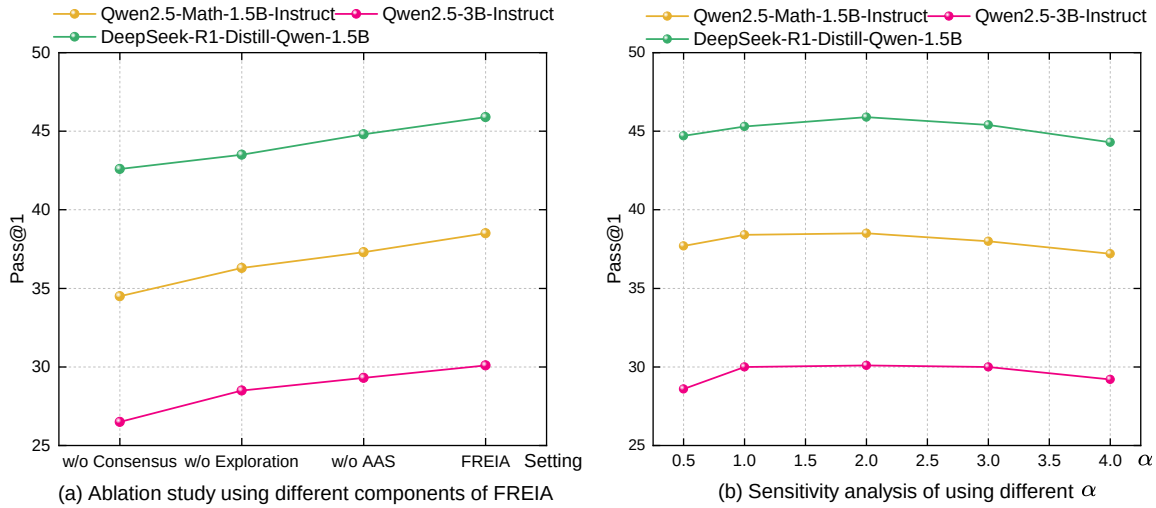
Remarkably, FREIA rivals or even surpasses the supervised GRPO. This performance gain is largely attributed to the granularity and density of the learning signal. While GRPO relies on binary feedback offering limited guidance, FER provides *continuous feedback from the trade-off between consensus and exploration*. Therefore, by encouraging diverse reasoning paths aligned with the model's internal knowledge, FREIA acquires robust patterns that generalize better to unseen complexities. This is further evidenced in Figure 5 and Table 5, where FREIA showcases strong transferability to SQL generation and multi-modal reasoning. As a result, minimizing free energy cultivates a more adaptive reasoning capability than outcome supervision.

4.2.2 Training Dynamics (RQ2)

To elucidate the mechanisms of FREIA, we analyze the evolution of key metrics during training.

Computational Efficiency. Figure 6 compares the total wall-clock training time using different methods. Despite incorporating FER and AAS, FREIA maintains training efficiency analogous to other baselines. This suggests that the overhead from FER and AAS is minimal, establishing FREIA as an efficient framework for unsupervised self-improvement. Further analysis of FREIA's computational efficiency is shown in Appendix C.7.

Free-Energy-Inspired Training Dynamics. Figure 7 offers insights into the learning process in FREIA. First, the consistent decrease in policy entropy (Figure 7(a)) and the upward trend in C_G (Figure 7(b)) indicate that the model gradually reduces internal uncertainty and concentrates on high-consensus answers. This is driven by the dynamic interplay between the two FER components. Initially, the Exploration term promotes the discovery of diverse reasoning paths. As the model's reasoning improves, the Consensus term exhibits an upward trend shown in Figure 7(d), reinforcing the discovered high-quality solutions. However, the Exploration term in Figure 7(c) maintains fluctuat-



(a) Ablation study using different components of FREIA

(b) Sensitivity analysis of using different α

Figure 8: Ablation study and hyperparameter sensitivity analysis of FREIA. (a) Average Pass@1 of the full FREIA compared to its ablated variants. (b) Average Pass@1 under varying values of the hyperparameter α.

ing throughout training. This confirms that FREIA sustains necessary exploration to mitigate early convergence. Additional analysis on output diversity is provided in Appendix C.6.

4.2.3 Ablation Study and Sensitivity Analysis (RQ3 & RQ4)

Impact of Key Components (RQ3). As shown in Figure 8(a), a consistent improvement is observed from ablated variants to FREIA, confirming that each innovation is essential. The most pronounced performance degradation occurs in the *w/o Consensus* variant, suggesting that consensus is a primary driver of self-improvement. The results in *w/o Exploration* variant also decline, highlighting the need of mitigating early convergence. While the *w/o AAS* variant outperforms other ablations, it remains inferior to the full FREIA. This gap demonstrates that static advantage shaping is inadequate for refining the policy optimization process.

Hyperparameter Sensitivity (RQ4). Figure 8(b) examines the impact of the parameter α, which controls the sensitivity of belief refinement within FER. Specifically, at lower α values, the system exhibits excessive stochasticity, resulting in unstable reasoning and difficulty in consolidating reliable paths. Conversely, excessively high α values strengthens consensus, leading to early convergence toward suboptimal solutions. Crucially, performance reaches its maximum at an intermediate α value, indicating that FREIA effectively balances the trade-off between exploration and consensus. Furthermore, it exhibites strong robustness, obviating the need for fine-grained hyperparameter tuning to achieve competitive performance. Further analysis on the rollout size G and additional results are shown in Appendices C.8 and D, respectively.

4.3 Case Study

To illustrate the advantages of FREIA over other unsupervised methods, we present a case study in Appendix C.9. Specifically, the problem involves a geometry problem with a strict inequality constraint. Notably, FREIA correctly outputs the solution by systematically enumerating all potential vertex configurations and filtering candidates against the constraint to isolate the valid answer. In contrast, TTRL falls into a reasoning shortcut by assuming a default vertex order, leading to a result that violates the explicit condition (*i.e.*, $x > 7$). Entropy and Intuitor fail to reason effectively, degenerating into repetitive loops without performing actual calculations. This demonstrates FREIA's superior capability in complex reasoning compared to other unsupervised baselines.

5 Conclusion

In this work, we propose FREIA, a unified framework that integrates unsupervised exploration and reasoning alignment under the Free Energy Principle. It introduces Free Energy-Driven Reward (FER) and Adaptive Advantage Shaping (AAS) without ground-truth supervision. Extensive experiments across multiple reasoning benchmarks showcase that FREIA promotes self-improvement and achieves competitive reasoning performance. This approach provides a new paradigm for alignment in scenarios where labeled data is unavailable.

Limitations

Although FREIA exhibits strong capabilities, we identify specific avenues for future enhancement. (1) Our current resource budget limited the empirical scope to models with a maximum of 3B parameters. Nevertheless, we anticipate that the principles underlying FREIA's mechanisms will scale effectively to larger foundation models. (2) The Group Confidence relies exclusively on final answer distributions, neglecting the semantic nuances of intermediate reasoning paths. Developing a more advanced mechanism incorporating process-level monitoring would enhance the model's efficacy. (3) AAS employs batch-level skewness as a proxy for the collective learning state, which may obscure intra-batch variations. A more granular approach adapting to heterogeneous learning dynamics within a single batch represents a valuable direction for future optimization.

Ethical Considerations

We are committed to responsible AI research and adhere to the following principles: (1) To guarantee data privacy and reproducibility, we exclusively utilized publicly available datasets for training and evaluation, thereby avoiding any risk to private user data. (2) The study relied entirely on open-source LLMs. While we recognize the environmental impact associated with the energy consumption required for training, we believe improving sample efficiency is a step toward greener AI.

References

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. 2025. The unreasonable effectiveness of entropy minimization in llm reasoning. *arXiv preprint arXiv:2505.15134*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Christopher L Buckley, Chang Sub Kim, Simon McGregor, and Anil K Seth. 2017. The free energy principle for action and perception: A mathematical review. *Journal of mathematical psychology*, 81:55–79.

Huayu Chen, Kaiwen Zheng, Qinsheng Zhang, Ganqu Cui, Yin Cui, Haotian Ye, Tsung-Yi Lin, Ming-Yu Liu, Jun Zhu, and Haoxiang Wang. 2025. Bridging supervised learning and reinforcement learning in math reasoning. *arXiv preprint arXiv:2505.18116*.

MAA Codeforces. American invitational mathematics examination-aime 2024, 2024.

Karl Friston. 2010. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138.

Karl J Friston, Thomas Parr, Yan Yufik, Noor Sajid, Catherine J Price, and Emma Holmes. 2020. Generative models, linguistic communication and active inference. *Neuroscience & Biobehavioral Reviews*, 118:42–64.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Guanhua Huang, Tingqiang Xu, Mingze Wang, Qi Yi, Xue Gong, Siheng Li, Ruibin Xiong, Kejiao Li, Yuhao Jiang, and Bo Zhou. 2025. Low-probability tokens sustain exploration in reinforcement learning with verifiable reward. *arXiv preprint arXiv:2510.03222*.

Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, et al. 2024. Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent ai. *Advances in Neural Information Processing Systems*, 37:19209–19253.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, et al. 2024a. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*, 13(9):9.

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, et al. 2024b. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *Advances in Neural Information Processing Systems*, 36.

Pengyi Li, Matvey Skripkin, Alexander Zubrey, Andrey Kuznetsov, and Ivan Oseledets. 2025. Confidence is all you need: Few-shot rl fine-tuning of language models. *arXiv preprint arXiv:2506.06395*.

Jia Liu, ChangYi He, YingQiao Lin, MingMin Yang, FeiYang Shen, and ShaoGuo Liu. 2025. Ettrl: Balancing exploration and exploitation in llm test-time reinforcement learning via entropy mechanism. *arXiv preprint arXiv:2508.11356*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6774–6786.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. 2025. Maximizing confidence alone improves reasoning. *arXiv preprint arXiv:2505.22660*.

Archiki Prasad, Weizhe Yuan, Richard Yuanzhe Pang, Jing Xu, Maryam Fazel-Zarandi, Mohit Bansal, Sainbayar Sukhbaatar, Jason Weston, and Jane Yu. 2024. Self-consistency preference optimization. *arXiv preprint arXiv:2411.04109*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Yikun Wang, Yibin Wang, Dianyi Wang, Zimian Peng, Qipeng Guo, Dacheng Tao, and Jiaqi Wang. 2025. Geometryzero: Improving geometry solving for llm with group contrastive policy optimization. *arXiv preprint arXiv:2506.07160*.

Xumeng Wen, Zihan Liu, Shun Zheng, Zhijian Xu, Shengyu Ye, Zhirong Wu, Xiao Liang, Yang Wang, Junjie Li, Ziming Miao, et al. 2025. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. 2024b. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921.

Wenzhen Yuan, Shengji Tang, Weihao Lin, Jiacheng Ruan, Ganqu Cui, Bo Zhang, Tao Chen, Ting Liu, Yuzhuo Fu, Peng Ye, and Lei Bai. 2025. Wisdom of the crowd: Reinforcement learning from coevolutionary collective feedback. *arXiv preprint arXiv:2508.12338*.

Yanzhi Zhang, Zhaoxi Zhang, Haoxiang Guan, Yilin Cheng, Yitong Duan, Chen Wang, Yue Wang, Shuxin Zheng, and Jiyan He. 2025a. No free lunch: Rethinking internal feedback for llm reasoning. *arXiv preprint arXiv:2506.17219*.

Zizhuo Zhang, Jianing Zhu, Xinmu Ge, Zihua Zhao, Zhanke Zhou, Xuan Li, Xiao Feng, Jiangchao Yao, and Bo Han. 2025b. Co-reward: Self-supervised reinforcement learning for large language model reasoning via contrastive agreement. *arXiv e-prints*, pages arXiv–2508.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. 2025. Learning to reason without external rewards. *arXiv preprint arXiv:2505.19590*.

Pan Zhou, Xingyu Xie, Zhouchen Lin, and Shuicheng Yan. 2024. Towards understanding convergence and generalization of adamw. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, et al. 2025. Ttrl: Test-time reinforcement learning. *arXiv preprint arXiv:2504.16084*.

A Appendix A: Additional Experimental Setup

A.1 Models and Benchmarks

Table 2 details the access links for the models and benchmarks utilized in this study. All datasets are publicly available under the CC BY-SA 4.0 license, which enables modifications and inclusion of additional annotations on the original datasets.

A.2 Answer Extraction and Equivalence

To ensure accurate reward computation, we applied task-specific strategies to extract answers and calculate unique answers.

Mathematical Reasoning and Geometry3K. We employed standard answer extraction techniques used in prior work (Shao et al., 2024). The final answer was extracted from the answer box and compared via exact string matching.

SQL Generation. Since different SQL queries can be semantically equivalent despite having different surface forms, we determined equivalence based on *SQL execution results* rather than string matching. To be specific, two SQL queries were considered the same answer if they yielded identical execution results on the given databases. The set of unique answers was defined by the set of distinct SQL execution outputs.

B Appendix B: Pseudo Code for FREIA

We provide the pseudo code for FREIA in Algorithm 1.

C Appendix C: Further Analysis

C.1 Theoretical Derivation: FER as Precision-Weighted Active Inference

While FER is conceptually grounded in the *Free Energy Principle* (FEP), it can also be expressed as a tractable approximation of the maximization of the *Precision-Weighted Expected Free Energy* (EFE). The correspondence between the EFE objective and the proposed reward formulation is formalized through two operational assumptions.

General Objective. In Active Inference, agents minimize the expected free energy $G(\pi)$. Maximizing $-G(\pi)$ can be decomposed into two components: identifying preferred outcomes (*Pragmatic Value*) and reducing uncertainty (*Epistemic Value*) (Friston et al., 2020):

$$-G(\pi) \approx \underbrace{\mathbb{E}_{Q(o)}[\ln P(o \mid \theta_{goal})]}_{\text{Pragmatic}} + \underbrace{\mathbb{E}_{Q(o)}[D_{KL}(Q(s \mid o) \parallel Q(s))]}_{\text{Epistemic}} \tag{11}$$

Here, o, θ_{goal}, and s denote the observed generated answers, the parameters of the goal distribution (*i.e.*, prior preferences), and the latent reasoning states, respectively.

To adapt this formulation to unsupervised RL, we define the *Precision-Weighted EFE Objective* using the precision $\beta \in [0, 1]$:

$$J(\pi, \beta) = \beta \cdot (\text{Pragmatic}) + (1 - \beta) \cdot (\text{Epistemic}) \tag{12}$$

Definition 1 (Empirical Precision)**.** *The precision β of the model's current belief state is defined as the complement of the normalized entropy of the generated sample distribution W:*

$$\beta \coloneqq C_G = \begin{cases} 1.0, & \text{if } M = 1, \\ 1 - \frac{H(W)}{\log M}, & \text{if } M > 1, \end{cases} \tag{13}$$

where M is the number of unique samples. This formulation provides a tractable estimator of the model's consensus confidence.

To make Eq. (12) computationally tractable without ground-truth supervision, we introduce the following two approximations:

Assumption 1 (Consensus-Truth Proxy)**.** *In the absence of ground-truth supervision, the mode of the current empirical distribution y^* is assumed to approximate the optimal goal state. The pragmatic utility is modeled as a Dirac delta function centered on the consensus:*

$$\begin{aligned} &P(o \mid \theta_{goal}) \approx \delta(o = y^*) \\ &\implies \textit{Utility}_{prag} = \mathbb{I}(o = y^*) \end{aligned} \tag{14}$$

Assumption 2 (Satiable Information Gain)**.** *The epistemic value is associated with the surprisal of the outcome $-\ln P(o)$, where $P(o)$ is approximated by w_i (Eq. (1)). Since raw surprisal is unbounded, the utility of information is assumed to follow a saturation curve confined to $[0, 1]$:*

$$\textit{Utility}_{epis} \approx \tanh(-\ln w_i) \tag{15}$$

Based on these assumptions, we now present the theorem that formally justifies the FER reward formulation.

Name	Access Link	# Samples
	Models	
Qwen2.5-Math-1.5B-Instruct	https://huggingface.co/Qwen/Qwen2.5-Math-1.5B-Instruct	-
Qwen2.5-3B-Instruct	https://huggingface.co/Qwen/Qwen2.5-3B-Instruct	-
DeepSeek-R1-Distill-Qwen-1.5B	https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B	-
Qwen2.5-Coder-3B-Instruct	https://huggingface.co/Qwen/Qwen2.5-Coder-3B-Instruct	-
Qwen2.5-VL-3B-Instruct	https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct	-
	Benchmarks	
MATH	https://huggingface.co/datasets/HuggingFaceH4/MATH	12000 training data
MATH500	https://huggingface.co/datasets/HuggingFaceH4/MATH-500	500 evaluation data
AIME24	https://huggingface.co/datasets/HuggingFaceH4/aime_2024	30 evaluation data
AIME25	https://huggingface.co/datasets/HuggingFaceH4/aime_2025	30 evaluation data
AMC23	https://huggingface.co/datasets/math-ai/amc23	40 evaluation data
Minerva	https://huggingface.co/datasets/svc-huggingface/minerva-math	272 evaluation data
OlympiadBench	https://huggingface.co/datasets/knoveleng/OlympiadBench	674 evaluation data
Spider	https://yale-lily.github.io/spider	1034 evaluation data
BIRD	https://bird-bench.github.io/	9428 and 1534 data for training and evaluation
Geometry3K	https://huggingface.co/datasets/hiyouga/geometry3k	2100 and 601 data for training and evaluation

Table 2: Details of models and benchmarks utilized in this study.

Theorem 1 (FER Derivation)**.** *Under Assumptions 1 and 2, maximizing the Precision-Weighted EFE objective $J(\pi, \beta)$ with empirical precision $\beta = C_G$ is equivalent to maximizing the expected FER reward R_{FER}.*

Proof. Let the empirical probability of an outcome o_i be w_i. Substituting into Eq. (12):

1. **Pragmatic Term:** By Assumption 1, the pragmatic component becomes the indicator function $\mathbb{I}(o_i = y^*)$, which corresponds to r_{cons}.

2. **Epistemic Term:** By Assumption 2, the epistemic component becomes $\tanh(-\ln w_i)$, corresponding to r_{explore}.

3. **Precision Modulation:** By Definition 1, the weighting factor β is given by the group confidence C_G.

Substituting these into J:

$$\begin{aligned} J(\pi, C_G) &= C_G \cdot \mathbb{E}[\mathbb{I}(o = y^*)] \\ &\quad + (1 - C_G) \cdot \mathbb{E}[\tanh(-\ln w_i)] \\ &= \mathbb{E}_\pi \Big[\underbrace{C_G \cdot r_{\text{cons}} + (1 - C_G) \cdot r_{\text{explore}}}_{R_{\text{FER}}} \Big] \end{aligned} \tag{16}$$

Therefore, the reward R_{FER} used in policy updates is a direct realization of the Precision-Weighted EFE objective. □

This derivation establishes FER as a concrete instantiation of Active Inference, in which the agent adaptively shifts from *risk minimization* (consensus) to *information seeking* (exploration) according to the statistical reliability (C_G) of its outputs. This theoretical connection provides a principled foundation for FER's adaptive balance between consensus and exploration in unsupervised reasoning.

Algorithm 1 Free Energy-Driven Reinforcement Learning with Adaptive Shaping (FREIA)

Input: Policy model π_θ, reference model $\pi_{\theta_{\text{ref}}}$, training dataset $\mathcal{D}$;
Hyperparameters: Learning rate η, KL coefficient β, group size G, temperature τ, sharpening factor α;
Output: Optimized policy model π_θ^*;

for step $t = 1$ to T **do**
  Sample query $q \sim \mathcal{D}$;
  /* — ***Step 1: Generation & Belief Formation*** — */
  Sample G responses $\mathcal{Y} = \{y_i\}_{i=1}^G \sim \pi_\theta(\cdot|q)$ and derive answers $\{a_i\}_{i=1}^G$;
  Identify M unique answers and compute frequencies f_j for $j \in \{1, \ldots, M\}$;
  Compute belief distribution $w_j \leftarrow \text{Softmax}(\alpha \cdot \log f_j)$; ▷ Eq. (1)
  /* — ***Step 2: Dual-Objective Reward Calculation*** — */
  Compute Group Confidence: $C_G \leftarrow \begin{cases} 1.0 & \text{if } M = 1 \\ 1 - \frac{H(W)}{\log M} = 1 - \frac{-\sum_j w_j \log w_j}{\log M} & \text{if } M > 1 \end{cases}$; ▷ Eq. (13)
  for $i = 1$ to G **do**
    $r_{\text{cons}}(y_i) \leftarrow \begin{cases} 1.0 & \text{if } a_i = \text{Vote}(\{a_k\}_{k=1}^G) \\ 0.0 & \text{otherwise} \end{cases}$;
    $r_{\text{explore}}(y_i) \leftarrow \tanh(-\log w_i)$;
    $R_i \leftarrow C_G \cdot r_{\text{cons}}(y_i) + (1 - C_G) \cdot r_{\text{explore}}(y_i)$; ▷ Eq. (5)
  end for
  /* — ***Step 3: Adaptive Advantage Shaping (AAS)*** — */
  Compute statistics: $\mu_R \leftarrow \text{mean}(\{R_i\})$, $\sigma_R \leftarrow \text{std}(\{R_i\})$;
  Calculate skewness: $S \leftarrow \frac{1}{G} \sum_{i=1}^G \left(\frac{R_i - \mu_R}{\sigma_R + \epsilon} \right)^3$; ▷ Eq. (7)
  Compute modulation weights: $w_{\text{pos}} \leftarrow \sigma(-S)$, $w_{\text{neg}} \leftarrow \sigma(S)$;
  for $i = 1$ to G **do**
    $\tilde{A}_i \leftarrow \frac{R_i - \mu_R}{\sigma_R + \epsilon}$;
    if $\tilde{A}_i > 0$ **then**
      $\hat{A}_i \leftarrow w_{\text{pos}} \cdot \tilde{A}_i$; ▷ Positive Advantage Modulation
    else
      $\hat{A}_i \leftarrow w_{\text{neg}} \cdot \tilde{A}_i$; ▷ Negative Advantage Modulation
    end if
  end for
  /* — ***Step 4: Policy Optimization*** — */
  $\mathcal{L}(\theta) \leftarrow \mathbb{E}\Big[\frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min\Big(r_{i,t}(\theta) \hat{A}_i,\ \text{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon) \hat{A}_i \Big) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \Big]$;
  Update parameters: $\theta \leftarrow \theta + \eta \nabla_\theta \mathcal{L}(\theta)$;
end for

C.2 The Necessity of Adaptive Precision via Group Confidence

To validate the efficacy of FEP, we conducted an ablation study where the Group Confidence (C_G) was replaced with static mixing coefficients.

C.2.1 Experimental Setup

In the standard FREIA framework, the final reward is computed as an adaptive convex combination:

$$R_{\text{total}} = C_G \cdot r_{\text{cons}} + (1 - C_G) \cdot r_{\text{explore}} \quad (17)$$

where $C_G \in [0, 1]$ represents a confidence score indicating the degree of group consensus for each sample. In this ablation, C_G was replaced with a fixed hyperparameter $\lambda \in \{0.2, 0.5, 0.8\}$ for all training samples:

$$R_{\text{fixed}}(\lambda) = \lambda \cdot r_{\text{cons}} + (1 - \lambda) \cdot r_{\text{explore}} \quad (18)$$

This setup forces a static trade-off, where $\lambda = 0.2$ and $\lambda = 0.8$ represent a high-exploration strategy and a high-consensus strategy, respectively.

C.2.2 Analysis of Static vs. Dynamic Strategies

Experimental results in Table 8 reveal a clear performance hierarchy and expose the weaknesses of static weighting:

Excessive Exploration ($\lambda = 0.2$). A low $\lambda = 0.2$ yielded the poorest average performance. By penalizing convergence toward consensus, this configuration injected noise into learning signals, even when the consensus was correct. Maximizing surprisal without adequate consensus consistently led to divergent behavior.

Blind Conformity ($\lambda = 0.8$). Greater reliance on consensus improved overall performance but introduced systematic risk. When the majority was wrong, a high λ reinforced erroneous reasoning. Without the ability to down-weight consensus under uncertainty, the model remained vulnerable to majority-driven bias. Therefore, a fixed λ implicitly assumes a constant level of uncertainty across

Setup: Sample Size $G = 8$. **Scenario:** A high-ambiguity task with a weak false consensus.
Generated Answers: $A_1 \sim A_5$ are incorrect (Weak Majority). $A_6 \sim A_8$ are correct (Strong Minority).
Empirical Frequencies: $f_{wrong} = 5/8 = 0.625$, $f_{correct} = 3/8 = 0.375$.

1. Belief Formation (Step 1, with $\alpha = 2.0$):
- Raw Proportions (f): $\{0.625, 0.375\}$.
- Squared Proportions (f^α): $\{0.625^2 \approx 0.39,\ 0.375^2 \approx 0.14\}$. Sum $\Sigma \approx 0.53$.
- Belief Weights ($w = f^\alpha/\Sigma$):
  $w_{wrong} = 0.39/0.53 \approx$ **0.74**, $w_{correct} = 0.14/0.53 \approx$ **0.26**.

2. Group Confidence (C_G) Calculation:
- Entropy $H(W) = -(0.74 \ln 0.74 + 0.26 \ln 0.26) \approx 0.57$.
- Max Entropy $\ln M = \ln 2 \approx 0.69$ (2 unique answer clusters).
- $C_G = 1 - (0.57/0.69) =$ **0.17** (Low confidence, high uncertainty).

3. Reward Component Calculation:
- **consensus (r_{cons}):** Mode is "Wrong". $r_{\text{exploit}}^{wrong} = 1.0$, $r_{\text{exploit}}^{correct} = 0.0$.
- **Exploration (r_{explore}):** Using $\tanh(-\ln w)$.
  $r_{\text{explore}}^{wrong} = \tanh(-\ln(0.74)) \approx$ **0.29** (Low information gain).
  $r_{\text{explore}}^{correct} = \tanh(-\ln(0.26)) \approx$ **0.87** (High information gain).

4. Final Reward Integration ($R = C_G \cdot r_{\text{cons}} + (1 - C_G) \cdot r_{\text{explore}}$):
- **Wrong Majority ($A_1 \sim A_5$):** $R_{wrong} = 0.17 \cdot 1.0 + 0.83 \cdot 0.29 \approx$ **0.41**.
- **Correct Minority ($A_6 \sim A_8$):** $R_{correct} = 0.17 \cdot 0.0 + 0.83 \cdot 0.87 \approx$ **0.72**.

5. Comparison with Standard Baseline (Majority Voting):
- **Standard Voting Reward:** $R_{wrong} = 1.0$ (Win), $R_{correct} = 0.0$ (Lose).

Conclusion: In standard voting, the model reinforces the incorrect majority ($R = 1.0$). In FREIA, the high conflict ($C_G = 0.17$) triggers a shift to exploration. Consequently, the correct minority receives a significantly higher reward than the majority (**0.72** vs **0.41**), successfully reversing the policy update direction towards the truth.

Table 3: A case study on False Consensus. FREIA detects the high uncertainty and inverts the reward signal, prioritizing the correct minority over the incorrect majority.

all problems, which is theoretically unsound.

Dynamic Adaptation (C_G). FREIA achieved the highest accuracy across almost all benchmarks by employing the adaptive C_G as an *intelligent gate*:

- **Low Uncertainty ($C_G \to 1$):** Under strong consensus, FREIA operated close to the $\lambda = 1.0$ case, fully exploiting the correct path.
- **High Uncertainty ($C_G \to 0$):** Under low consensus, FREIA automatically lowered the weight of r_{cons} and amplified r_{explore}, enabling exploration of alternative reasoning paths rather than overfitting to unreliable majority answers.

C.3 Theoretical Analysis of AAS

In this section, an analytical framework is provided to interpret the stability mechanisms in AAS. The gradient dynamics are analyzed under stylized regimes of reward skewness. This analysis illustrates that AAS functions as a variance-reduction mechanism in practice.

The behavior of AAS is examined from two complementary perspectives: *Influence Limitation* for outlier robustness in positive-skew regimes, and *Variance Damping* for stability in negative-skew regimes.

Proposition 1 (Gradient Stabilization in Skewed Regimes)**.** *Under extreme skewness, AAS analytically reduces the gradient magnitudes of outliers in positive-skew regimes and mitigates variance amplification in negative-skew regimes.*

Proof. We analyze two typical scenarios of training dynamics:

Scenario 1: Positive Skew. Consider a batch with sparse high rewards (*e.g.*, a single high-reward answer among many low-reward answers). The distribution exhibits large positive skew $S > 0$.

(1) **Standard Normalization Instability:** Consider a batch of size G where only one sample y_{rare} receives a large reward R_{rare}, while others have near-zero rewards. Using the standard normalization $\tilde{R}_i = \frac{R_i - \mu_R}{\sigma_R}$, the batch mean is $\mu_R \approx$

R_{rare}/G, and the variance is dominated by this single outlier:

$$\sigma_R^2 \approx \frac{(R_{\text{rare}} - \mu_R)^2}{G} \approx \frac{R_{\text{rare}}^2}{G} \Rightarrow \sigma_R \approx \frac{R_{\text{rare}}}{\sqrt{G}} \tag{19}$$

Therefore, the normalized reward of the rare sample becomes:

$$\tilde{R}_{\text{rare}} \approx \frac{R_{\text{rare}}}{R_{\text{rare}}/\sqrt{G}} = \sqrt{G} \tag{20}$$

This yields $\tilde{A}_{\text{rare}} \approx \sqrt{G}$, making the gradient $\nabla J \propto \sqrt{G} \cdot \nabla \ln \pi(y_{\text{rare}})$ overly dependent on one sample. If the outlier is wrong, it triggers a destructive high-variance policy update.

(2) **AAS Damping:** Since $S > 0$, AAS applies a scaling weight $w_{\text{pos}} = \sigma(-S) < 1$, giving an effective advantage $\hat{A}_{\text{rare}} = w_{\text{pos}} \cdot \tilde{A}_{\text{rare}} < \tilde{A}_{\text{rare}}$. This bounds the outlier's influence and keeps the policy update conservative in unsupervised settings.

Scenario 2: Negative Skew. Consider a regime where low-value rewards are rare. The batch contains predominantly high-reward answers ($r \approx 1$) with probability $1 - p$, and occasional low-reward answers ($r \approx 0$) with probability $p \ll 1$. The skewness $S < 0$.

(1) **Variance Sensitivity in Standard RL:** The standard deviation of rewards is $\sigma \approx \sqrt{p}$. The normalized advantage for the rare failure is:

$$\tilde{A}_{fail} = \frac{0 - (1-p)}{\sqrt{p}} \approx -\frac{1}{\sqrt{p}} \tag{21}$$

As $p \to 0$ (*i.e.*, the model approaches perfection), the magnitude $|\tilde{A}_{fail}| \to \infty$. This results in gradient magnitudes inversely proportional to $\sqrt{p}$, which can introduce high variance and destabilize policy optimization.

(2) **AAS Variance Damping:** In the negatively skewed regime, the distribution of rewards is dominated by high-reward samples. AAS attenuates these negative outliers through a sigmoid-based weighting $w_{neg} = \sigma(S)$. As S decreases, w_{neg} rapidly approaches zero, effectively damping the contribution of extreme negative advantages:

$$\lim_{S \to -\infty} w_{neg} \cdot A_{fail} = 0 \tag{22}$$

As a result, this alleviates gradient explosion and promotes stable training. □

C.4 The Dual Role of Belief Sharpening: Exploration Control and Noise Filtering

In Section 4.1, the belief-sharpening mechanism was defined as $w_j \propto f_j^{\alpha}$, where α acts as a structural parameter controlling the learning dynamics. This section investigates the influence of α from two perspectives: (i) its isolated effect on r_{explore}, and (ii) its theoretical grounding within Generalized Bayesian Inference and the bias–variance trade-off.

Isolation of Exploration Dynamics. Let f_{win} denote the frequency of the dominant consensus answer. The exponent α defines the reward landscape and modulates the incentive for policy deviation:

(1) $\alpha = 1$ **(Raw Surprisal):** This case corresponds to using the raw frequency distribution without sharpening ($w_{\text{win}} = f_{\text{win}}$). During early training or under challenging conditions, consensus is weak, producing a flat distribution and a high exploration reward for the dominant answer. This results in redundant exploration, where the model repeatedly examines known solutions rather than searching for novel reasoning trajectories.

(2) **High α (Mode Collapse):** As $\alpha \to \infty$, $w_{\text{win}} \to 1$ and $r_{\text{explore}}(w_{\text{win}}) \to 0$. While this reduces redundant exploration, it prematurely suppresses diversity. If the initial majority is incorrect, gradients associated with the correct minority vanish, causing convergence to a local optimum.

Empirical results show that $\alpha = 2$ achieves superior performance, indicating that the sharpened reward $\tanh(-\log w_j)$ offers a more informative signal than the raw form $\tanh(-\log f_j)$ for guiding exploration.

Principled Selection: Bayesian and Bias-Variance Perspectives. The role of α can be further grounded in the *Power Posterior* framework of Generalized Bayesian Inference:

$$P(s|D) \propto [P(D|s)]^{\eta} \cdot P(s) \tag{23}$$

In FREIA, the observed frequency f_j serves as the empirical likelihood $P(D|s)$, and the computed weight w_j represents the posterior belief. The parameter α corresponds to the exponent η, functioning as a *confidence temperature*.

- **Noise Filtering (Variance Reduction):** In unsupervised reinforcement learning, raw frequencies ($\alpha = 1$) are highly variable and noise-sensitive. Increasing α acts as a denoising mechanism, suppressing the long tail of

stochastic errors and concentrating probability mass around the mode. This adjustment reduces the variance of the learning signal and stabilizes gradient estimation.

- **Gradient Preservation (Bias Control):** In contrast to extreme hard-filtering ($\alpha \to \infty$), the setting $\alpha = 2$ retains a small but non-zero bias. This configuration maintains adaptability to adjust incorrect consensus, ensuring that plausible minority solutions remain learnable despite imperfections in current consensus.

As a result, α functions as a controllable temperature parameter balancing exploration and noise suppression.

C.5 Empirical Robustness of Skewness Estimation

To validate the stability of AAS, we tracked the evolution of the skewness throughout training for each batch, as shown in Figure 9. The trajectories reveal two critical insights regarding stability:

- **Temporal Stability:** All models display a coherent downward trend with minimal high-frequency fluctuations. This smooth evolution indicates that the estimation variance is negligible, implying that the skewness metric remains stable across training iterations.

- **Direction Persistence:** Throughout the entire training process, the skewness consistently stays within the negative domain. This persistent negativity provides a substantial margin from zero, confirming that the direction of AAS modulation is reliably determined by the intrinsic reward landscape. It also demonstrates statistical immunity to stochastic variations arising from individual samples.

C.6 Empirical Analysis of Output Diversity

We evaluated the generative diversity of FREIA by comparing it with TTRL and GRPO on MATH500. Three complementary metrics were employed to characterize diversity across semantic, lexical, and logical dimensions.

(1) **Average Cosine Similarity (ACS):** ACS measures the semantic homogeneity among generated solutions. We computed cosine similarity between sentence embeddings[2] for all response pairs given the same question. A lower ACS value indicates greater semantic variance.

[2] https://huggingface.co/Qwen/Qwen3-Embedding-0.6B

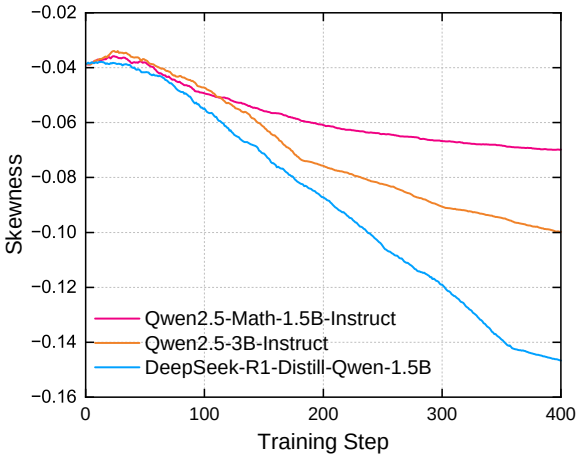


Figure 9: The evolution of the reward skewness throughout training.

(2) **Self-BLEU:** This metric quantifies lexical overlap by calculating the average BLEU-4 score of each generated solution against the remaining samples in a batch. Lower Self-BLEU scores correspond to reduced verbatim repetition.

(3) **GPT-Judged Diversity Score:** We further assessed conceptual diversity using GPT-4o as an external judge, which rated the distinctness of reasoning paths on a $0 \sim 3$ scale. Higher scores (3) were assigned to fundamentally different mathematical approaches, while lower scores (0) indicated superficial rephrasing.

Method	ACS (↓)	Self-BLEU (↓)	GPT-Diversity (↑)
GRPO	0.88	0.79	1.05
TTRL	0.76	0.68	1.23
FREIA	**0.62**	**0.51**	**1.56**

Table 4: Diversity metrics on the MATH500 dataset using Qwen2.5-Math-1.5B-Instruct. ↓ indicates lower is better; ↑ indicates higher is better.

As shown in Table 4, GRPO attained the highest ACS and Self-BLEU scores, confirming its susceptibility to severe mode collapse. While TTRL alleviated this issue via consistency voting, it remained constrained by the dominance of the consensus term, limiting its ability to maintain high diversity. In contrast, FREIA achieved the lowest similarity scores across all metrics. These results demonstrated that the exploration reward term r_{explore} effectively penalized redundancy in low-confidence states, guiding the policy toward distinct reasoning trajectories rather than repetitive consensus. The

observed improvements align with FEP introduced in Appendix C.1.

GPT-Judge Prompt. The prompt used for diversity scoring is given as follows:

"Given the following set of correct solutions to a math problem, evaluate their diversity. Assign a score from 0 to 3: 0 = Almost identical wording; 1 = Different wording but same logic; 2 = Slightly different logical steps; 3 = Fundamentally different mathematical approaches. Output only the single numerical score."

C.7 Computational Efficiency Analysis

This section analyzes the training efficiency of FREIA through the lens of time complexity.

In RL-based frameworks, the computational bottleneck is dominated by the LLM's forward pass (reasoning generation) and backward pass (gradient optimization). Let N, L, and G denote the number of model parameters, sequence length, and group size (rollouts per prompt), respectively. The complexity of these backbone operations scales as $\mathcal{O}(G \cdot L \cdot N)$.

In contrast, the mechanisms introduced by FREIA operate solely on scalar reward values, bypassing high-dimensional tensor manipulations:

- C_G **Calculation:** This involves sorting the reward distribution of size G, with a complexity of $\mathcal{O}(G \log G)$.
- **Skewness and AAS Weighting:** This requires computing statistical moments of the reward distribution, which is linear in group size (*i.e.*, $\mathcal{O}(G)$).

Given that G is typically small (*e.g.*, $G = 8$) while N is in the billions, the inequality $\mathcal{O}(G \cdot L \cdot N) \gg \mathcal{O}(G \log G)$ holds in practice across different model scales. Therefore, the additional computational overhead introduced by FREIA is theoretically negligible.

C.8 Sensitivity Analysis: The Impact of Group Size G

The group size G is a critical hyperparameter in FREIA, governing the fidelity of statistical estimates (*i.e.*, skewness, C_G) and the computational cost. While larger groups provide more robust gradient signals, they also increase memory and time consumption. To determine the optimal trade-off, we conducted an ablation study using Qwen2.5-Math-1.5B-Instruct with group sizes $G \in \{4, 6, 8\}$.

As shown in Table 9, performance is positively correlated with G, confirming that larger populations stabilize skewness-based advantage estimation. However, a distinct trend of *diminishing performance gains* is also observed. Since rollout generation costs scale linearly with G, scaling beyond $G = 8$ yields insufficient gains to justify the added computational burden. Therefore, $G = 8$ offered a balanced configuration that maintains statistical robustness while minimizing unnecessary computation.

C.9 Case study

Case Study: Reasoning Process Comparison Using Different Unsupervised Methods

Original Question:
The coordinates of a parallelogram are $(5, 3)$, $(6, 8)$, $(7, 4)$ and (x, y) and $x > 7$. What is the value of $x + y$?

FREIA Answer (Correct): ✓
To find the coordinates of the fourth vertex (x, y), we must consider that the vertices can be connected in different orders to form a parallelogram. Let the given points be $A(5, 3)$, $B(6, 8)$, and $C(7, 4)$. There are three possible positions for the fourth vertex $D(x, y)$, depending on which pair of points forms a diagonal.

Case 1: AB **is a diagonal.**
The midpoint of AB must equal the midpoint of CD.

$$\frac{5+6}{2} = \frac{7+x}{2} \implies 11 = 7 + x \implies x = 4$$
$$\frac{3+8}{2} = \frac{4+y}{2} \implies 11 = 4 + y \implies y = 7$$

Here $x = 4$, which is not greater than 7. **This case is invalid.**

Case 2: AC **is a diagonal.**
The midpoint of AC must equal the midpoint of BD.

$$\frac{5+7}{2} = \frac{6+x}{2} \implies 12 = 6 + x \implies x = 6$$
$$\frac{3+4}{2} = \frac{8+y}{2} \implies 7 = 8 + y \implies y = -1$$

Here $x = 6$, which is not greater than 7. **This case is invalid.**

Case 3: BC **is a diagonal.**
The midpoint of BC must equal the midpoint of AD.

$$\frac{6+7}{2} = \frac{5+x}{2} \implies 13 = 5 + x \implies x = 8$$
$$\frac{8+4}{2} = \frac{3+y}{2} \implies 12 = 3 + y \implies y = 9$$

Here $x = 8$, which satisfies the condition $x > 7$.

Thus, the valid coordinates are $(8, 9)$.
The value of $x + y = 8 + 9 = 17$.
The final answer is $\boxed{17}$**.**

TTRL Answer (Incorrect – Logic Error): ✗
Let the vertices of the parallelogram be $A(5, 3)$, $B(6, 8)$, $C(7, 4)$ and $D(x, y)$ in counterclockwise order. In a parallelogram, the diagonals bisect each other. This means the midpoint of AC is the same as the midpoint of BD.
Midpoint of AC:

$$x_m = \frac{5+7}{2} = 6, \quad y_m = \frac{3+4}{2} = 3.5$$

Now set this equal to the midpoint of BD:

$$6 = \frac{6+x}{2} \implies 12 = 6 + x \implies x = 6$$
$$3.5 = \frac{8+y}{2} \implies 7 = 8 + y \implies y = -1$$

Method	Geometry3K	Spider	BIRD	Avg.
Base	25.7	70.2	43.6	46.5
GRPO	$35.6_{\pm 1.1}$	$73.8_{\pm 0.5}$	$55.7_{\pm 0.9}$	$55.0_{\pm 0.8}$
TTRL	$35.3_{\pm 0.8}$	$74.0_{\pm 0.4}$	$56.0_{\pm 0.2}$	$55.1_{\pm 0.4}$
Intuitor	$34.8_{\pm 0.9}$	$72.8_{\pm 0.6}$	$54.5_{\pm 0.8}$	$54.0_{\pm 0.8}$
Entropy	$34.4_{\pm 0.7}$	$72.5_{\pm 0.5}$	$54.0_{\pm 0.7}$	$53.6_{\pm 0.6}$
FREIA	$\mathbf{36.1}_{\pm 0.5}$	$\mathbf{74.4}_{\pm 0.2}$	$\mathbf{56.4}_{\pm 0.3}$	$\mathbf{55.6}_{\pm 0.3}$

Table 5: Generalization performance on SQL generation (Spider, BIRD) and multi-modal reasoning (Geometry3K). The results are reported as mean and standard deviation across 3 random seeds (format: $\text{Mean}_{\pm \text{Std}}$).

So the point is $(6, -1)$. The question asks for $x + y$.

$$x + y = 6 + (-1) = 5$$

The final answer is $\boxed{5}$.

(Note: TTRL fails because it defaults to the standard vertex ordering assumption and ignores the explicit $x > 7$ constraint, a common failure mode in consensus-based methods where the majority of rollouts miss the subtle constraint.)

Entropy Answer (Collapse – Repetition Loop): ✗
To solve this problem, we use the property that the diagonals of a parallelogram bisect each other.
Let the points be $A(5, 3)$, $B(6, 8)$, $C(7, 4)$. We need to find $D(x, y)$.
The midpoint of the diagonal is the average of the coordinates.
The midpoint of the diagonal is the average of the coordinates.
The midpoint of the diagonal is the average of the coordinates.
The midpoint of the diagonal is the average of the coordinates.
The midpoint of the diagonal is the average of the coordinates.
The midpoint of the diagonal is the average of the coordinates.
The midpoint of the diagonal is the average of the coordinates...

(Note: The model enters a degenerate loop, repeating the same phrase endlessly, a typical collapse symptom in entropy minimization.)

Intuitor Answer (Collapse – Degenerate Reasoning): ✗
Step 1: Identify the coordinates. The coordinates are $(5, 3)$, $(6, 8)$, $(7, 4)$ and (x, y).
Step 2: Use the formula. We use the formula for the parallelogram.
Step 1: Identify the coordinates. The coordinates are $(5, 3)$, $(6, 8)$, $(7, 4)$ and (x, y).
Step 2: Use the formula. We use the formula for the parallelogram.
Step 1: Identify the coordinates. The coordinates are $(5, 3)$, $(6, 8)$, $(7, 4)$ and (x, y).
Step 2: Use the formula...

(Note: Similar to Entropy, Intuitor (Self-Certainty) collapses into a high-confidence repetition loop without performing actual calculation.)

D Appendix D: Additional Experimental Results

Model	Method	MATH500	AIME24	AIME25	AMC23	Minerva	Olympiad	Avg.
Qwen2.5-Math-1.5B	w/o Consensus	74.4	10.0	6.7	47.5	29.4	38.9	34.5
	w/o Exploration	74.8	13.3	10.0	50.0	30.1	39.5	36.3
	w/o AAS	75.2	13.3	13.3	50.0	31.3	40.5	37.3
	FREIA	**75.4**	**13.3**	**16.7**	**52.5**	**32.0**	**41.2**	**38.5**
Qwen2.5-3B-Instruct	w/o Consensus	63.2	3.3	3.3	35.0	24.6	29.5	26.5
	w/o Exploration	64.2	6.7	6.7	37.5	25.4	30.7	28.5
	w/o AAS	64.8	6.7	6.7	**40.0**	26.1	31.5	29.3
	FREIA	**65.2**	**10.0**	**10.0**	37.5	**26.1**	**31.9**	**30.1**
DeepSeek-R1-Distill	w/o Consensus	81.2	16.7	16.7	65.0	29.0	47.2	42.6
	w/o Exploration	81.6	16.7	16.7	67.5	29.8	48.5	43.5
	w/o AAS	82.0	20.0	16.7	70.0	30.5	**49.6**	44.8
	FREIA	**82.2**	**20.0**	**20.0**	**72.5**	**31.3**	49.4	**45.9**

Table 6: Detailed Ablation Study (RQ3). We compare FREIA against removing key components: w/o Consensus, w/o Exploration, and w/o AAS.

Model	Coefficient	MATH500	AIME24	AIME25	AMC23	Minerva	Olympiad	Avg.
Qwen2.5-Math-1.5B	$\alpha = 0.5$	74.8	13.3	16.7	50.0	31.3	40.1	37.7
	$\alpha = 1.0$	75.2	13.3	16.7	52.5	31.6	40.8	38.4
	$\alpha = 2.0$	**75.4**	**13.3**	**16.7**	**52.5**	**32.0**	**41.2**	**38.5**
	$\alpha = 3.0$	75.0	13.3	16.7	50.0	32.0	40.9	38.0
	$\alpha = 4.0$	74.8	13.3	13.3	50.0	31.3	40.2	37.2
Qwen2.5-3B-Instruct	$\alpha = 0.5$	64.6	6.7	6.7	37.5	25.4	30.6	28.6
	$\alpha = 1.0$	65.0	10.0	10.0	37.5	25.7	31.8	30.0
	$\alpha = 2.0$	**65.2**	**10.0**	**10.0**	37.5	**26.1**	**31.9**	**30.1**
	$\alpha = 3.0$	64.8	10.0	6.7	40.0	25.7	31.9	30.0
	$\alpha = 4.0$	64.6	10.0	6.7	37.5	25.0	31.5	29.2
DeepSeek-R1-Distill	$\alpha = 0.5$	81.6	16.7	20.0	70.0	30.5	49.3	44.7
	$\alpha = 1.0$	82.0	16.7	20.0	72.5	30.9	49.9	45.3
	$\alpha = 2.0$	**82.2**	**20.0**	**20.0**	**72.5**	**31.3**	**49.4**	**45.9**
	$\alpha = 3.0$	81.8	20.0	16.7	72.5	31.3	50.1	45.4
	$\alpha = 4.0$	81.6	16.7	16.7	70.0	30.9	49.6	44.3

Table 7: Hyperparameter Sensitivity Analysis (RQ4). Impact of the coefficient α on model performance across diverse benchmarks.

Setting	Mixing Strategy	MATH500	AIME24	AIME25	AMC23	Minerva	Olympiad	Avg.
Fixed Mixing	$\lambda = 0.2$	74.6	10.0	13.3	48.5	30.5	39.8	36.1
Fixed Mixing	$\lambda = 0.5$	74.8	13.3	13.3	50.0	30.8	40.2	37.1
Fixed Mixing	$\lambda = 0.8$	75.0	13.3	16.7	50.0	31.3	40.7	37.8
FREIA	**Dynamic** C_G	**75.4**	**13.3**	**16.7**	**52.5**	**32.0**	**41.2**	**38.5**

Table 8: Ablation study on the mixing coefficient λ using Qwen2.5-Math-1.5B-Instruct. Baselines with fixed λ represent a static trade-off between Consensus (λ) and Surprisal ($1 - \lambda$). FREIA uses Group Confidence (C_G) to adaptively modulate this trade-off per instance.

Setting	MATH500	AIME24	AIME25	AMC23	Minerva	Olympiad	Avg.	Growth Rate
FREIA ($G = 4$)	74.6	13.3	13.3	47.5	30.5	40.1	36.6	-
FREIA ($G = 6$)	75.2	13.3	16.7	50.0	31.6	41.1	38.0	+1.4%
FREIA ($G = 8$)	**75.4**	**13.3**	**16.7**	**52.5**	**32.0**	**41.2**	**38.5**	+0.5%

Table 9: Ablation study on Group Size G using Qwen2.5-Math-1.5B-Instruct. We report the Pass@1 score across all datasets. Growth Rate indicates the performance gain relative to the previous group size setting.