Diagnosing Unsupervised RL for Self-Compression in Mathematical Reasoning

Anonymous ACL submission

Abstract

Current approaches to reasoning compression on Large Language Models (LLMs) rely on external supervision or crafted prompts, leaving the models' intrinsic compression capabilities under unsupervised settings unexplored. Concurrently, unsupervised reinforcement learning (RL) has shown that intrinsic rewards shape reasoning behaviors. If intrinsic rewards can shape general reasoning, we hypothesize that applying intrinsic constraints as rewards can also stimulate a model's capability for *self-compression*: the pruning of reasoning steps driven by unsupervised signals. To verify this, we establish two probing methods and design a controlled framework that subjects models with varying mathematical priors to these unsupervised constraints, mapping their training dynamics into a measurable semantic space. We find that models undergo compression before policy collapse, and their resistance to this collapse is proportional to their foundational reasoning capability. By tracking the manifold volume of the explored semantic space, we provide a metric to diagnose training stability boundaries.

1 Introduction

Inference-time scaling advances the mathematical reasoning capabilities of LLMs. Methods such as outcome-based RL, including Group Relative Policy Optimization (GRPO; Shao et al., 2024), improve performance by allowing models to generate extensive chains of thought. However, this growth in reasoning length results in "overthinking" (Sui et al., 2025). Verbose and redundant reasoning steps incur computational costs, increase inference latency, and degrade model performance. Consequently, reasoning compression is a requirement for practical deployment.

Current compression approaches rely on external supervision or crafted prompts (Han et al., 2025; Xia et al., 2025; Team et al., 2025), leaving the intrinsic compression capabilities of models under unsupervised settings unexplored. Concurrently, unsupervised RL optimizes models using intrinsic rewards such as entropy minimization (Zhao et al., 2026; Prabhudesai et al., 2025; Agarwal et al., 2026) or self-consistency (Zhang et al., 2025; Zuo et al., 2025) without ground-truth labels. These methods demonstrate that intrinsic rewards shape a model's reasoning behaviors. Intuitively, if intrinsic rewards shape general reasoning, applying such intrinsic constraints could also serve as signals for driving reasoning compression. To explore this, we establish two probing methods and hypothesize that applying them as rewards in RL stimulates and diagnoses a model's capability for self-compression.

To verify this hypothesis, we design the controlled evaluation framework illustrated in Figure 1. The framework begins with a scene that models capable of reasoning under an extensive response length budget (e.g., 32K tokens) are forced to operate under a limited budget (e.g., 4K tokens) by avoiding generating abundant tokens. To adapt to this scene, we define *self-compression* as optimizing the models via unsupervised intrinsic rewards. To understand the mechanisms of this process, we decompose our analysis into three research questions. We evaluate the capability-driven performance outcomes (RQ1), reduce the dimensionality of the training dynamics to extract independent semantic axes (RQ2), and diagnose the stability boundaries of the compression process (RQ3).

RQ1: Do foundational reasoning capabilities dictate self-compression? To investigate this, we apply two unsupervised rewards, Shannon Entropy (Ent) and Length Penalty (LP), representing an entropy-weighted length constraint and a uniform verbosity penalty, as probing constraints via GRPO to models across a spectrum of mathematical priors and monitor their performance under length budgets. We find that although the global optima

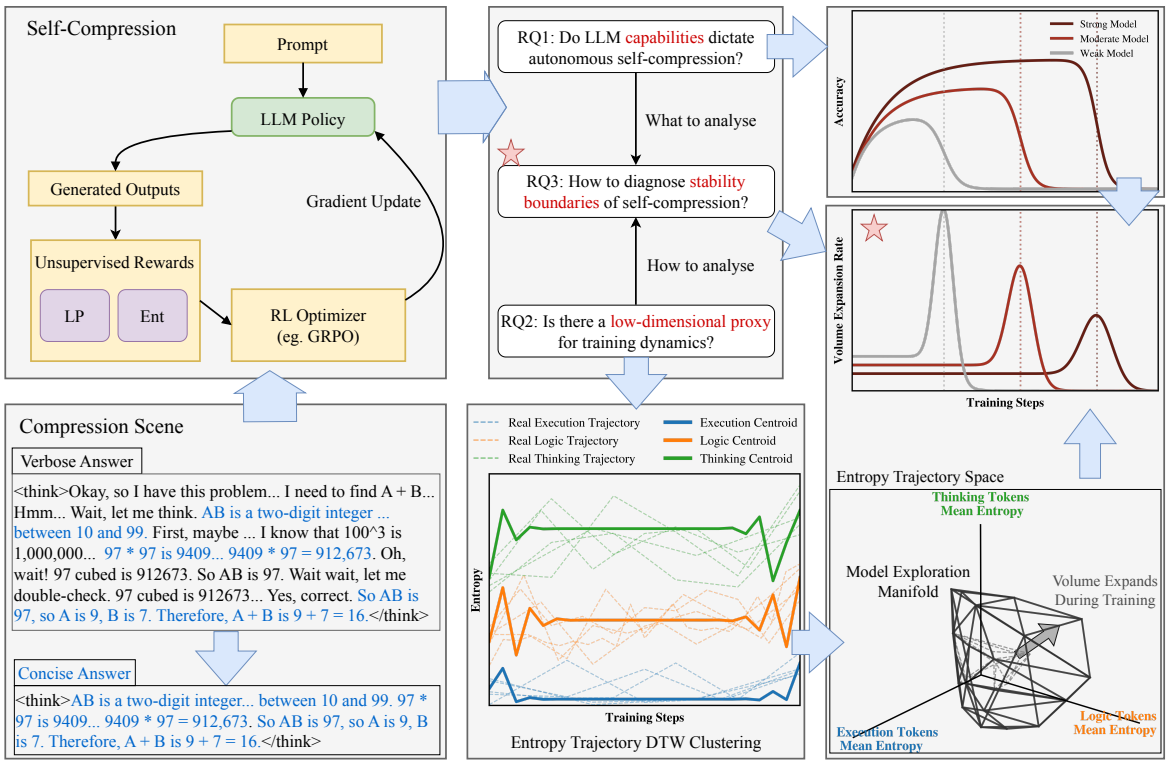


Figure 1: **Research Framework. (a) RQ1:** Effect of base model capabilities on self-compression. **(b) RQ2:** 3D projection of training dynamics using DTW clustering. **(c) RQ3:** Explaining accuracy collapse through semantic manifold expansion.

for these rewards are degenerate distributions, the models undergo compression before collapsing. A model's resistance to this collapse reflects its capacity for self-compression and is proportional to its foundational reasoning capability.

RQ2: Is there a low-dimensional proxy to interpret training dynamics? Recent studies reveal that the training trajectories of deep networks explore a low-dimensional manifold (Mao et al., 2024). To understand the mechanisms driving capability-driven compression, we analyze the temporal dynamics of token-level entropy throughout the training process using time-series clustering. We discover a structural consistency. Across varying model capabilities and reward formulations, the reasoning trajectories stratify into three semantic states: Execution, Logic, and Thinking. This stratification allows us to project the training dynamics into a measurable, low-dimensional semantic space.

RQ3: How can we mathematically diagnose the stability boundaries of self-compression? Operating within this measurable semantic space, we quantify the bounds of the models' exploration by tracking the convex hull volume of their reasoning trajectories across training steps and correlating its expansion rate with validation accuracy. This analysis establishes that policy collapse synchronizes with a spike in the manifold expansion rate. High-capacity models resist this topological explosion. The volume expansion rate serves as a metric to diagnose training stability boundaries.

In summary, our findings validate our hypothesis: models possess a capability for self-compression stimulated by unsupervised constraints. This compression is a transient state bounded by the model's reasoning prior, rather than a stable convergence. By mapping this process into a semantic space, our work provides a geometric explanation for policy collapse during unsupervised RL and a metric for designing future compression algorithms.

2 Related Work

Reasoning Compression via Supervised Methods and Prompting Current approaches to reasoning compression on LLMs involve prompt engineering, SFT, and distillation. Prompting strategies instruct models to reduce verbosity or to meet a token budget (Han et al., 2025). SFT methods use pruned reasoning traces for model training to filter

out redundant steps in generation (Han et al., 2025; Xia et al., 2025). Distillation transfers reasoning logic from teacher models to smaller chunks (Chen et al., 2025) or continuous latent spaces (Shen et al., 2025). These approaches require external supervision, prompts, or teacher models. Our work explores the intrinsic compression capabilities of models. Models prune redundant reasoning under unsupervised constraints, without external labeled data.

Length Constraints in Supervised RL Supervised RL (S-RL) frameworks integrate length penalties into verifiable reward functions to penalize long trajectories. Examples include length penalties (Team et al., 2025) and budget forcing strategies under the Information Bottleneck principle (Massoli et al., 2026). These studies combine length constraints with verifiable ground-truth rewards. Our approach removes the verifiable reward. We use length or entropy constraints as the intrinsic objective. This design evaluates the boundary of model stability during compression.

Unsupervised RL Unsupervised RL optimizes models using intrinsic rewards, including entropy minimization (Zhao et al., 2026; Prabhudesai et al., 2025; Agarwal et al., 2026) or self-consistency (Zhang et al., 2025; Zuo et al., 2025). These methods improve reasoning accuracy without ground-truth labels. Our work connects unsupervised RL and reasoning compression. We frame intrinsic rewards as structural diagnostic tools. The temporal entropy dynamics map into a 3D semantic space. We provide a geometric metric, Manifold Expansion Rate, to identify the collapse boundary during the self-compression process.

3 Training Methods

To effectively translate our unsupervised intrinsic constraints into policy updates, we employ GRPO.

3.1 Policy Optimization

For a given mathematical problem query q, we sample a group of G response token trajectories $\{o_1, o_2, \ldots, o_G\}$ from the old policy $\pi_{\theta_{old}}$. For each trajectory o_i, we compute an intrinsic scalar reward r_i, the specific formulations of which are introduced in Section 3.2.

The advantage A_i for each rollout is then calculated by standardizing the rewards within the sampled group:

$$A_i = \frac{r_i - \mu_{\mathbf{r}}}{\sigma_{\mathbf{r}}} \tag{1}$$

where $\mu_{\mathbf{r}}$ and $\sigma_{\mathbf{r}}$ are the mean and standard deviation of the group rewards $\mathbf{r} = \{r_1, \ldots, r_G\}$, respectively. The policy model π_θ is then optimized by maximizing the objective function of GRPO:

$$\mathcal{J}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\rho_i(\theta)A_i, \text{clip}(\epsilon)A_i\right)\right)\right] \tag{2}$$

where $\rho_i(\theta) = \pi_\theta(o_i|q)/\pi_{\theta_{old}}(o_i|q)$ is the importance sampling ratio, ϵ is the clipping hyperparameter to prevent destructively large policy updates, and π_{ref} is the reference model. Note that we explicitly exclude the KL divergence term $\mathbb{D}_{KL}$ across all settings. While typically used to anchor the policy to a reference model and prevent over-optimization (Schulman et al., 2017), our primary objective is to diagnose the intrinsic boundary of stability. Introducing a KL penalty would act as an artificial safety net that functionally overlaps with our unsupervised constraints analysis.

3.2 Rewards

To formulate the intrinsic reward r_i used in our GRPO framework, we evaluate two specific unsupervised constraints: LP and Ent. Specifically, LP represents a direct and uniform penalty on verbosity, mirroring the explicit length constraints often used in recent Supervised RL papers. Ent, on the other hand, acts as an entropy-weighted length penalty. Inspired by the recent successes of unsupervised Entropy Minimization, this formulation encourages concise responses dynamically by penalizing tokens with high uncertainty.

During the RL training phase, for a generated reasoning response of length T, let $p_t(v)$ denote the predicted probability of token v from the vocabulary at generation step t.

LP. We establish LP as a pure length punishment:

$$R_{\text{LP}} = -T/T_{\max} \tag{3}$$

where $T_{\max}$ denotes the maximum response length.

Ent. The reward employs the Shannon entropy:

$$R_{\text{Ent}} = -\sum_{t=1}^{T}\mathcal{H}(p_t) = \sum_{t=1}^{T}\sum_{v\in\mathcal{V}} p_t(v)\log(p_t(v)) \tag{4}$$

Model	AIME 26 (Pass@1)
Llama-3.1-8B-Instruct	3.3
DeepSeek-R1-Distill-Llama-8B	33.3
Qwen3-8B	**63.3**

Table 1: **Foundational Reasoning Capability.** Pass@1 accuracy on the AIME 26 dataset evaluated with an extended 32K response budget.

Because token entropy $\mathcal{H}(p_t)$ is strictly non-negative, the cumulative sum monotonically decreases as the sequence elongates. The model is pressured to become certain and concise at the same time.

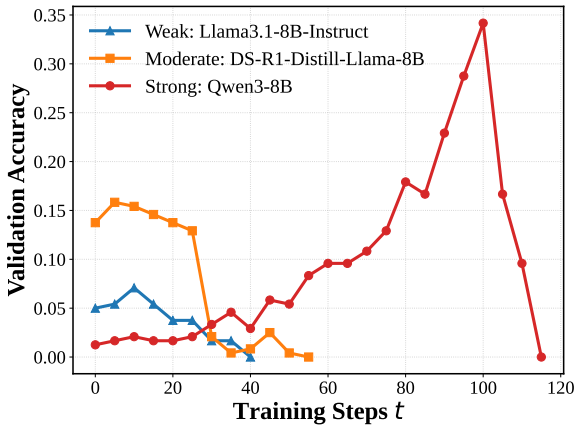
S-RL. S-RL optimizes the policy using verifiable ground-truth correctness. This configuration provides a reference for a stable training state, allowing us to evaluate the boundaries of autonomous self-compression when such supervision is removed.

4 Experimental Settings

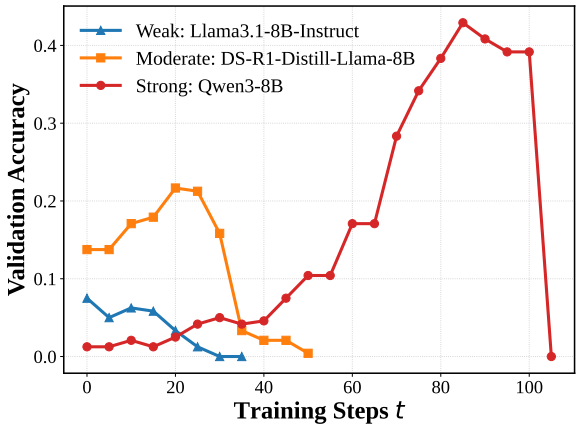
The primary objective of our experiments is to demonstrate how a model's foundational reasoning capability dictates its capacity for autonomous self-compression. Further hyperparameter details are provided in Appendix A.

Models. We evaluate four models representing a spectrum of inherent reasoning capacities: Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen3-1.7B, Qwen3-8B (Yang et al., 2025), and DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI, 2025). When provided with a 32K response budget, their foundational reasoning performance on AIME26 (Zhang and Math-AI, 2026) is detailed in Table 1. This evaluation confirms a capability hierarchy from the weaker Llama-3.1-8B to the highly capable Qwen3-8B. All models use their native 'think' mode if available, or prompted Chain-of-Thought (CoT) to maximize reasoning.

Training and Evaluation Setup. The models are trained on DAPO-Math-17K (Yu et al., 2025) under a 4K token response budget. During training, we track validation performance on the AIME24 dataset. Because the AIME24 (Zhang and Math-AI, 2024) dataset is relatively small, using a single greedy decoding pass can lead to highly fluctuating results. To ensure a stable evaluation, we generate 8 independent responses per problem and report the expected Pass@1 accuracy. We evaluate the final generalization performance across standardized benchmarks: MATH500 (Hendrycks et al., 2021), Minerva Math (Lewkowycz et al., 2022), OlympiadBench (He et al., 2024), AIME24, AMC23 [1] and AIME26 (out-of-distribution).



(a) Method: Ent



(b) Method: LP

Figure 2: **Validation accuracy curves on 4K.** The vertical axis represents the validation accuracy evaluated on the AIME 24 dataset, reported as the expected Pass@1 (averaged across 8 independent generations per problem).

5 RQ1: Capabilities Dictate Self-Compression

To empirically trace the dynamics of self-compression and policy collapse, we first visualize the validation accuracy trajectories (expected Pass@1) across training steps under the Ent and LP formulations, as shown in Figure 2. The curves clearly depict a capability-driven spectrum of survival: the strongest model (Qwen3-8B) sustains reasoning performance and achieves significant ac-

[1] https://maa.org/math-competitions/

Model	Method	MATH-500			Olympiad			Minerva			AIME 24			AMC 23			AIME 26			AVG		
		P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10
Qwen3-1.7B	Base	70.2	81.4	84.9	30.1	41.6	46.9	30.6	42.0	46.0	6.0	14.2	19.1	47.2	61.4	68.0	7.1	15.5	20.0	42.7	54.0	58.5
	S-RL	78.8	89.5	91.8	39.5	52.0	56.0	**35.4**	**49.2**	**53.9**	17.9	31.0	37.8	56.6	76.0	82.1	12.7	26.0	31.1	51.0	63.3	67.0
	LP	**80.0**	**90.4**	**92.4**	**42.6**	**55.3**	**59.1**	34.4	47.9	52.4	20.0	34.2	39.3	57.8	79.7	86.8	**14.4**	**28.0**	**33.6**	**52.7**	65.0	68.5
	Ent	79.8	**90.4**	**92.4**	41.4	54.9	58.8	34.4	48.3	52.9	**22.5**	**42.6**	**49.3**	**61.3**	**81.8**	**88.1**	14.0	27.9	31.6	52.2	**65.1**	**68.7**
Qwen3-8B	Base	67.9	80.5	84.4	28.4	41.2	46.7	29.8	39.9	44.8	7.7	19.9	23.1	41.7	64.1	73.6	9.2	16.6	22.1	41.0	53.4	58.3
	S-RL	88.4	93.8	95.0	51.7	60.1	63.4	46.0	55.7	58.3	36.2	52.7	**60.4**	76.2	87.9	89.1	36.2	48.8	50.0	62.6	70.6	73.1
	LP	**88.7**	**94.2**	**95.2**	**52.3**	**61.1**	**64.2**	**46.1**	**56.2**	**59.1**	**39.2**	**55.8**	59.1	**78.8**	**93.0**	**94.9**	**36.7**	**51.1**	54.1	**63.1**	**71.5**	**73.8**
	Ent	86.6	93.1	94.5	49.9	58.2	61.0	43.9	54.4	58.2	31.7	48.0	53.5	74.1	89.4	91.5	34.2	49.8	**54.6**	60.7	69.3	71.9
DeepSeek-Distill Llama-8B	Base	69.4	87.1	90.8	35.7	52.9	58.8	22.7	40.2	47.8	16.5	30.3	37.9	50.8	80.6	86.9	14.4	27.4	32.7	43.9	61.5	67.0
	S-RL	**72.7**	**90.2**	**93.7**	**39.3**	**56.9**	**62.4**	**24.4**	**42.7**	**49.9**	**24.4**	**44.9**	**54.9**	**59.4**	**86.9**	**91.1**	**22.5**	**40.0**	**44.6**	**47.4**	**65.4**	**70.6**
	LP	70.5	88.5	92.5	36.6	53.9	59.5	22.2	40.7	48.3	21.7	43.1	52.8	50.8	81.4	87.0	17.5	33.8	39.4	44.7	62.9	68.4
	Ent	70.0	88.1	91.9	36.4	54.6	60.6	21.8	39.9	47.2	18.1	37.9	47.5	53.4	83.8	88.5	17.3	31.2	36.1	44.4	62.8	68.4
Llama3.1-8B Instruct	Base	45.5	**69.4**	**77.2**	13.2	29.8	38.2	21.4	39.8	47.1	5.2	14.3	20.8	22.7	**51.6**	**66.4**	0.4	2.1	4.2	24.9	44.1	52.1
	S-RL	**46.3**	69.0	76.4	**14.1**	**31.0**	38.8	25.7	**45.5**	**52.3**	4.6	13.2	18.7	23.6	46.8	56.1	**1.2**	3.1	3.3	**26.4**	**45.4**	52.7
	LP	*Collapse* (×)																				
	Ent	45.8	**69.4**	77.0	13.9	30.6	**38.9**	**25.8**	43.7	50.1	**6.5**	**16.7**	**23.3**	**24.1**	46.9	56.9	1.0	**4.9**	**9.2**	26.2	45.1	**52.8**

Table 2: **Peak Results on Mathematical Reasoning Benchmarks on 4K. Bold** denotes the overall best performance, and underlined indicates the best result among unsupervised methods within each model. $P@k$ means Pass@k.

curacy gains before eventually decaying, whereas the moderate model (DeepSeek-R1-Distill-Llama-8B) exhibits a lower, earlier peak. In stark contrast, the weakest model (Llama3.1-8B-Instruct) suffers from almost immediate catastrophic collapse under LP and fails to register meaningful improvements under Ent. These validation dynamics confirm that a model's intrinsic resistance to policy collapse and its potential for autonomous self-compression are strictly bounded by its foundational reasoning prior.

These trajectories reveal a transient state of self-compression. To prevent evaluation bias from a single dataset, we extract the checkpoint at the peak of each validation curve for broad testing across a suite of unseen mathematical benchmarks. As shown in Table 2, the generalization results align with the validation trajectories. For high-capacity models like the Qwen series, unsupervised methods generalize to broader datasets and surpass the supervised baseline (S-RL) in certain cases (e.g., Qwen3-8B achieves an average accuracy of 73.8% under LP compared to 73.1% under S-RL). Under the tight 4K length budget, S-RL tends to degenerate into a formatting constraint, making intrinsic rewards a more effective approach for actual logic compression. For moderate and weak models, the test performance mirrors their foundational reasoning priors. This evaluation confirms that the transient peaks represent a genuine compression of reasoning capabilities.

6 RQ2: Entropy Dynamics as a Low-Dimensional Proxy

Due to space limitations, we focus on Qwen3-8B and the Ent method for analysis in this section. Results for other models and reward formulations are provided in Appendix B.

Instead of analyzing static entropy values, we focus on temporal dynamics: the evolution of token entropies throughout the training process. We apply Soft-DTW (Cuturi and Blondel, 2018) to measure trajectory distance and Time-series K-means to group them. Formalizations of this algorithm are in Appendix B.1.

6.1 The Choice of K

To determine the parameter K, we evaluate DTW Inertia and the Silhouette Score. The DTW Inertia is the sum of squared distances between each sample and its cluster centroid:

$$\text{Inertia} = \sum_{k=1}^{K} \sum_{x_i \in C_k} \text{DTW}(x_i, \mu_k)^2 \qquad (5)$$

where C_k represents the k-th cluster and μ_k is its centroid. The Silhouette Score S evaluates separation and cohesion:

$$S = \frac{1}{N} \sum_{i=1}^{N} \frac{b_i - a_i}{\max(a_i, b_i)} \qquad (6)$$

where a_i is the mean intra-cluster DTW distance and b_i is the mean distance to the nearest neighboring cluster.

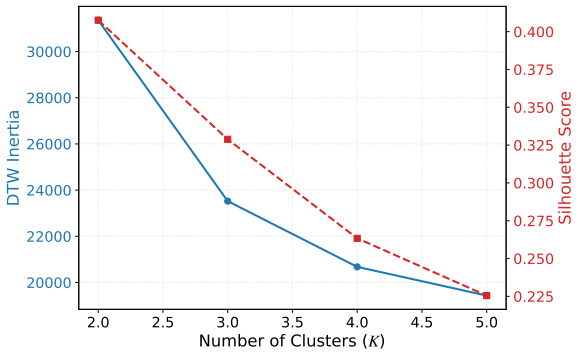


Figure 3: **Evaluation of Clustering K (Qwen3-8B-Ent).** The DTW inertia exhibits a distinct elbow at $K = 3$.

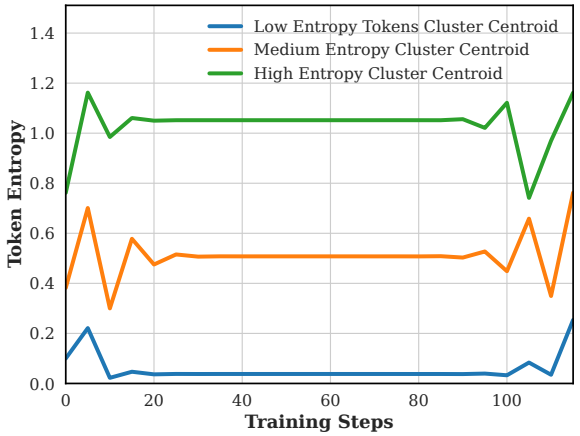


Figure 4: **Clustering Centroids of Token Entropy (Qwen3-8B-Ent).** Time-series DTW clustering reveals three clusters: Low, Medium, and High.

As shown in Figure 3 and Appendix B.2, $K = 2$ yields the highest Silhouette score. Clustering continuous time-series trajectories introduces a specific dynamic: increasing K separates continuous segments into adjacent clusters, reducing the inter-cluster distance (b_i) and the Silhouette score. The DTW inertia curve exhibits an elbow at $K = 3$. We select $K = 3$ to reduce intra-cluster variance while maintaining cluster separation.

Applying Time-Series K-means stratifies the entropy trajectories into low, medium, and high entropy clusters (Figure 4 and Appendix B.3). To evaluate this 3-cluster system, we measure the Overlap Similarity of the assigned tokens across different methods compared to the S-RL baseline:

$$\text{Overlap}(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)} \tag{7}$$

As illustrated in Figure 5 and Appendix B.4, the token overlap across training methods remains above

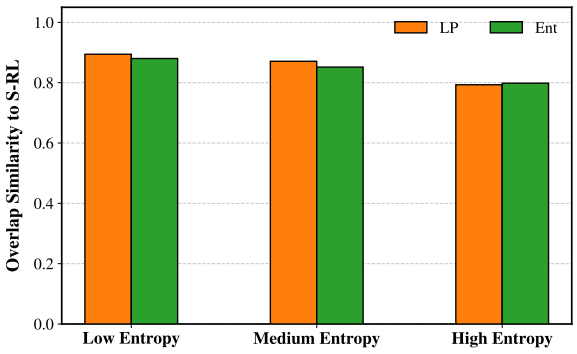


Figure 5: **Overlap Similarity of Clustered Tokens (Qwen3-8B).** Comparison of token overlap between the S-RL baseline and unsupervised rewards across the three entropy levels at $K = 3$.

Semantic Cluster (Entropy) & Behavior	Representative Tokens
Execution (Low) *Math Computation & Formatting*	**Iconic:** ˆ, \, $, k **Core:** =, $$, *, (,)
Logic (Mid) *Structural Transition & Check*	**Iconic:** `Now`, `check`, `Similarly` **Core:** `this`, `Let`, `for`
Thinking (High) *Hypothesis & Divergent Search*	**Iconic:** `analyze`, `contrad`, `Hmm` **Core:** `perhaps`, `maybe`, `suppose`

Table 3: **Token-Level Semantic Mapping (Qwen3-8B-Ent). Shape A (Iconic)** extracts the closest instance to the centroid. **Shape B (Core)** identifies the most frequent tokens in the closest centroid neighborhood.

0.8. This overlap indicates that the 3-cluster structure captures the model's reasoning process independent of the specific reward formulation.

6.2 Bridging Trajectories and Semantics

Table 3 and Appendix B.5 map these trajectory-based clusters to semantic profiles. The low, medium, and high entropy clusters correspond to Execution, Logic, and Thinking behaviors during mathematical reasoning. Profiling tokens via temporal entropy dynamics connects raw training trajectories with model semantics.

7 RQ3: Diagnosing Stability via Volume Expansion

By treating the entropy values of the three distinct semantic states as independent coordinate axes, we can project the opaque training dynamics into a measurable 3D semantic space. To quantify the model's exploration boundaries over time, we calculate the Convex Hull volume of the token trajectories at each training step.

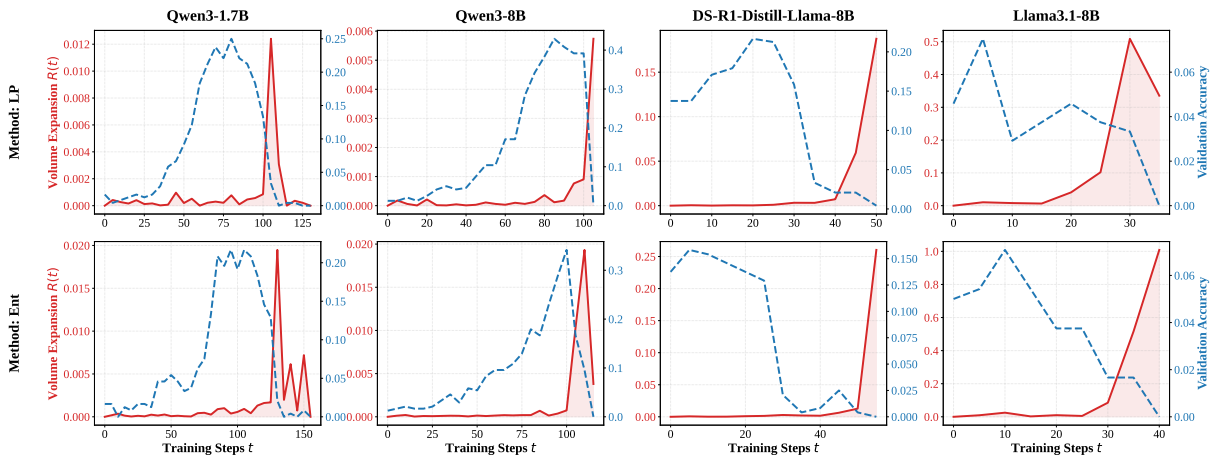


Figure 6: **Alignment of Manifold Expansion Rate and Accuracy Collapse.** The drop in validation accuracy (blue dashed line) is consistently accompanied by a sharp spike in the volume expansion rate $R(t)$ (red solid line).

7.1 Mapping Responses to a 3D Semantic Space

To construct this 3D semantic space, we evaluate the model every 5 training steps following the validation setting described in Section 4. Due to the computational cost of generating multiple responses across frequent evaluation intervals, we restrict this trajectory analysis to the AIME24 dataset.

During each evaluation step t, the model generates multiple responses. For any given response r, we first classify its generated tokens into the three semantic clusters defined earlier: Execution, Logic, and Thinking. Specifically, each token is assigned to a cluster based on the global token-to-cluster mapping derived in Section 6. Let C_k represent one of these clusters, where $k \in \{\text{exec}, \text{logic}, \text{think}\}$. We then calculate the average entropy of the tokens belonging to each cluster within this specific response (assigning $E_k(r) = 0$ if T_k is empty):

$$E_k(r) = \frac{1}{|T_k|} \sum_{x \in T_k} H(x) \tag{8}$$

where T_k is the set of tokens in response r that belong to cluster C_k, and $H(x)$ is the predicted entropy for token x. This operation reduces the complex sequence of tokens into a single coordinate point in our 3D semantic space:

$$\mathbf{p}_r = (E_{\text{exec}}(r), E_{\text{logic}}(r), E_{\text{think}}(r)) \tag{9}$$

Through this method, every response generated at step t is mapped to a specific coordinate. To measure how the model's exploration space expands over time, we collect all response coordinates generated from the beginning of training (step 0) up to the current step t. Let S_t denote this cumulative set of coordinates:

$$S_t = \bigcup_{\tau=0}^{t} \{\mathbf{p}_r \mid r \text{ at step } \tau\} \tag{10}$$

Finally, we quantify the model's exploration boundary at step t by calculating the volume of the Convex Hull formed by the coordinate set S_t:

$$V(t) = \text{Volume}(\text{ConvexHull}(S_t)) \tag{11}$$

We define this geometric boundary, represented by the Convex Hull, as the model's **semantic manifold**. Its cumulative volume $V(t)$ provides a mathematical representation of the total semantic space explored by the reasoning trajectories up to that point in training. The expansion rate is defined as the discrete derivative of the convex hull volume with respect to the training steps:

$$R(t) = \frac{V(t) - V(t - \Delta t)}{\Delta t} \tag{12}$$

where $\Delta t = 5$ in our experimental setup. For the initial evaluation step, the expansion rate is naturally defined as $R(0) = 0$.

7.2 Correlation Between Volume Expansion and Accuracy Collapse

To investigate the relationship between manifold topology and model performance, we visualize the expansion rate $R(t)$ alongside the validation accuracy. As depicted in Figure 6, the collapse of validation accuracy synchronizes with a spike in the expansion rate $R(t)$.

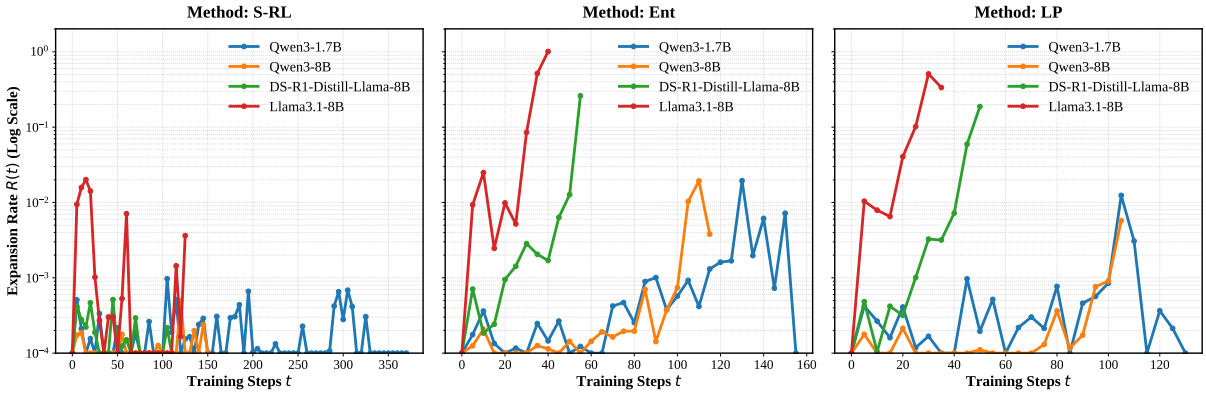


Figure 7: **Cross-Model Comparison of Manifold Expansion Rate $R(t)$.** The **logarithmic** scale highlights early-stage topological deviations. **(Left)** The S-RL method establish the baseline for a healthy state. **(Middle & Right)** Under the Ent and LP formulations, the timing of the volume explosion correlates with model capability.

This synchronization provides a geometric interpretation of policy degradation. As a model over-optimizes and loses logical reasoning capabilities, its internal semantic state loses structural constraints. This loss of bounds leads to the volume explosion defined in (11). After the complete collapse of accuracy, $R(t)$ returns to zero. A fully collapsed model degenerates into producing zero-length responses or repeating a single token. These degenerate trajectories contribute no new volume to the exploration manifold, resulting in a dormant expansion rate.

These observations establish volume expansion as a diagnostic indicator of model stability. This framework provides a lens to analyze the impact of model capacity. High-capacity models maintain a bounded manifold and sustain reasoning performance for longer durations, whereas weaker models suffer from rapid topological explosions. We delve into this capacity-driven resilience in the subsequent analysis.

7.3 Diagnosing Capability-Driven Topological Resilience

The synchronization between volume expansion and accuracy collapse enables the use of the expansion rate $R(t)$ as a diagnostic tool. We apply this metric to analyze the capability-driven performance disparities observed in Section 5.

Figure 7 plots the expansion rate $R(t)$ across varying model capacities and methods on a logarithmic scale. This scale isolates early-stage topological deviations prior to the volume spike. Under the supervised S-RL baseline (Figure 7, Left), $R(t)$ remains suppressed across all model scales. This bounded expansion defines a stable training state. When supervised targets regularize reasoning trajectories, the semantic manifold remains constrained regardless of the underlying model capacity. Under unsupervised rewards, the topology of the semantic manifold exposes the model's foundational reasoning capacity. Weaker models exhibit an early loss of structural constraints. Their $R(t)$ displays an exponential spike within the first 20 to 40 training steps.

Models with stronger mathematical priors restrict the volume of their semantic manifold for a longer duration. They sustain a flat $R(t)$ for over 100 steps before collapse. During the collapse phase, weaker models produce a steep gradient in the expansion curve, while stronger models produce a gradual ascent. High-capacity models resist the degradation of logical constraints and maintain a restricted volume under unsupervised signals.

8 Conclusion

We find that a model's capacity for self-compression is strictly bounded by its foundational reasoning prior. To diagnose the boundary of catastrophic collapse, we utilized temporal entropy dynamics to map reasoning trajectories into three semantic states within a 3D semantic space. Our analysis reveals that the manifold volume expansion rate synchronizes with the drop in validation accuracy, establishing this rate as a direct diagnostic metric for intrinsic structural resilience. By bridging raw training dynamics with interpretable semantics, this framework provides a practical basis for future LLM post-training, offering quantitative metrics for topology-aware regularization and compression algorithm designs.

Limitations

Scale. Due to the substantial computational overhead required by unsupervised RL, our evaluation is currently restricted to base models up to 8B parameters.

Application in S-RL. The autonomous self-compression phenomenon discussed here is established strictly as an intrinsic property of the model when subjected to unguided constraints. This work does not explore how these internal topological dynamics interact with external supervised data: whether the capability-driven resilience observed in unsupervised settings can be actively leveraged to enhance S-RL.

Theoretical Gaps in Training Dynamics. Our DTW clustering shows that the entropy centroids of different semantic states (Execution, Logic, Thinking) remain stable throughout the entire training process. We hypothesize this stability is an intrinsic structural bias inherited from the base model's pre-training, but a formal optimization proof is left for future work.

Diagnostic Scope of the Expansion Metric. The current framework functions as an observational diagnostic tool. Our study identifies the correlation between volume expansion and model collapse. We exclude the implementation of active interventions or regularization techniques to penalize volume acceleration. Utilizing the calculated expansion rate as an unsupervised stopping signal or a topology-aware regularization target in post-training requires subsequent validation.

Sensitivity and Temporal Delay of the Volume Metric. The volume expansion rate exhibits a temporal delay relative to the decline in validation accuracy. This delay originates from the geometric properties of the three-dimensional semantic space. A policy collapse along a single semantic dimension yields a restricted impact on the total convex hull volume. The manifold volume expands upon a near-complete policy collapse across all dimensions. Consequently, the expansion rate functions as a lagging indicator of global structural degradation. It lacks the mathematical sensitivity to detect the early, localized phases of accuracy decline.

References

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. 2026. The unreasonable effectiveness of entropy minimization in LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Xiaoshu Chen, Sihang Zhou, Ke Liang, Xiaoyu Sun, and Xinwang Liu. 2025. Skip-thinking: Chunk-wise chain-of-thought distillation enable smaller language models to reason better and faster. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12142–12157, Suzhou, China. Association for Computational Linguistics.

Marco Cuturi and Mathieu Blondel. 2018. Soft-dtw: a differentiable loss function for time-series. *Preprint*, arXiv:1703.01541.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. 2025. Token-budget-aware LLM reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24842–24855, Vienna, Austria. Association for Computational Linguistics.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *Preprint*, arXiv:2402.14008.

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. *Preprint*, arXiv:2103.03874.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models. *Preprint*, arXiv:2206.14858.

Jialin Mao, Itay Griniasty, Han Kheng Teoh, Rahul Ramesh, Rubing Yang, Mark K. Transtrum, James P. Sethna, and Pratik Chaudhari. 2024. The training

process of many deep networks explores the same low-dimensional manifold. *Proceedings of the National Academy of Sciences*, 121(12):e2310002121.

Fabio Valerio Massoli, Andrey Kuzmin, and Arash Behboodi. 2026. Reasoning as compression: Unifying budget forcing via the conditional information bottleneck. *Preprint*, arXiv:2603.08462.

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

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

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

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. 2025. CODI: Compressing chain-of-thought into continuous space via self-distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 677–693, Suzhou, China. Association for Computational Linguistics.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *Preprint*, arXiv:2503.16419.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, and 77 others. 2025. Kimi k1.5: Scaling reinforcement learning with llms. *Preprint*, arXiv:2501.12599.

Heming Xia, Chak Tou Leong, Wenjie Wang, Yongqi Li, and Wenjie Li. 2025. TokenSkip: Controllable chain-of-thought compression in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3351–3363, Suzhou, China. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. 2025. Right question is already half the answer: Fully unsupervised LLM reasoning incentivization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yifan Zhang and Team Math-AI. 2024. American invitational mathematics examination (aime) 2024.

Yifan Zhang and Team Math-AI. 2026. American invitational mathematics examination (aime) 2026.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. 2026. Learning to reason without external rewards. In *The Fourteenth International Conference on Learning Representations*.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. 2025. TTRL: Test-time reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

A Experiments' Settings

The detailed hyperparameters for our unsupervised RL training are provided in Appendix Table 4. Additionally, our generation hyperparameters for the validation phase roughly follow the recommended settings from the Qwen3 Technical Report (Yang et al., 2025).

B Clustering

B.1 Detailed Algorithm for Trajectory Construction

We formalize the data collection, temporal anchoring, and clustering process in Algorithm 1.

The core challenge is that the generated response for the same prompt p changes across training steps t (e.g., a model might output "Hmm, 5" at step 10, but simply "5" at step 50). To solve this, we use the tuple (p, y) as a unique semantic anchor, mapping the discrete token y back to its prompt p. This allows us to track the exact temporal evolution of the token's entropy regardless of when it appears in the generation sequence, ultimately forming the trajectory $\mathcal{T}_{p,y}$ for DTW clustering.

Algorithm 1 Token-Level Entropy Trajectory Construction and Semantic Clustering

Require: Validation dataset $\mathcal{D}_{val}$, Set of training evaluation steps $K = \{t_1, t_2, \ldots, t_M\}$
Require: Target number of clusters $C = 3$ (Execution, Logic, Thinking)
Ensure: Semantic clusters $\mathcal{C}$ and their corresponding centroids

Initialize an empty hash map $\mathcal{T}$ to store trajectories, keyed by (p, y).
for each evaluation step $t \in K$ **do**
    Load model checkpoint $\mathcal{M}_t$
    for each prompt $p \in \mathcal{D}_{val}$ **do**
        Generate response sequence $R_{p,t}$ using $\mathcal{M}_t(p)$
        Extract response tokens $Y_{p,t} = [y_1, y_2, \ldots, y_N]$
        Extract token-level entropies $H_{p,t} = [h_1, h_2, \ldots, h_N]$
        for $i = 1$ **to** N **do**
            $\mathcal{T}[(p, y_i)]$.append_record(step $= t$, entropy $= h_i$)
        end for
    end for
end for
$\mathcal{T}_{valid} \leftarrow \emptyset$
for each unique anchor (p, y) **in** $\mathcal{T}$ **do**
    Extract raw entropy sequence $\mathcal{S} = [h_{t_a}, \ldots, h_{t_b}]$ from $\mathcal{T}[(p, y)]$
    if length$(\mathcal{S}) \geq 2$ **then** ▷ Filter isolated tokens without temporal dynamics
        Normalize time axis of $\mathcal{S}$ to $\hat{t} \in [0, 1]$ based on the effective convergence point
        $\mathcal{T}_{valid} \leftarrow \mathcal{T}_{valid} \cup \{\mathcal{S}\}$
    end if
end for
Initialize Time-Series K-Means model $\mathcal{K}_{\text{DTW}}$ with $k = C$ and metric = Soft-DTW
$\mathcal{C}$, Centroids $\leftarrow \mathcal{K}_{\text{DTW}}$.fit_predict$(\mathcal{T}_{valid})$
return $\mathcal{C}$

Hyperparameter	Value
Optimization	
Actor Learning Rate	1×10^{-6}
Learning Rate Warmup Steps	10
Weight Decay	0.1
Mini-batch Size	24
Prompt Batch Size	8
Reinforcement Learning	
Responses per Prompt (K)	16
PPO Clip Ratio (ϵ)	0.2
Loss Aggregation Mode	seq-mean-token-mean
Max Prompt Length	1024
Generation (Rollout & Validation)	
Training Temperature	0.6
Training Top-p	1.0
Training Top-k	-1
Validation Top-p	0.95
Validation Top-k	20

Table 4: Hyperparameters for the RL trainings.

Data Construction. We track the generation process on the validation set every 5 training steps. A trajectory is defined as a sequence of entropy values anchored to a specific (prompt, token) pair. Let $\mathcal{T}_{i,j} = [h_{t_1,j}, h_{t_2,j}, \ldots, h_{t_M,j}]$ denote the entropy trajectory for the i-th token in j-th prompt's response. The endpoint of training corresponds to the onset of the accuracy plateau for supervised baselines, or the point of collapse (0% accuracy) for unsupervised runs.

Clustering Algorithm. We employ **Time-Series K-means** with a Soft-DTW distance measurement to cluster these trajectories. Unlike Euclidean distance, DTW aligns sequences that may vary in phase or speed, allowing us to group tokens based on the *shape* of their entropy evolution (e.g., "rising" vs. "converging") rather than their absolute values at specific steps.

B.2 Comprehensive K-Validity Evaluation Across All Models and Methods

To rigorously justify this hyperparameter choice and demonstrate its universal applicability, we conducted a comprehensive K-validity evaluation across all evaluated models and reward formulations.

As illustrated in Figure 8, across almost all experimental settings, we observe a consistent pattern: the Silhouette Score typically maintains a strong local optimum at $K = 3$, while the DTW Inertia exhibits a noticeable "elbow" effect. This widespread structural consistency empirically validates our choice of $K = 3$ as the optimal topological boundary for mapping the models' internal reasoning manifolds, regardless of their foundational capability or the applied reward constraint.

B.3 Visualization of Entropy-Driven Trajectory Clusters

Figure 9 presents the combined trajectory plots across all four models and reward formulations (S-RL, Ent, and LP). The results demonstrate the emergence of the Low, Mid, and High-Entropy states across all training settings.

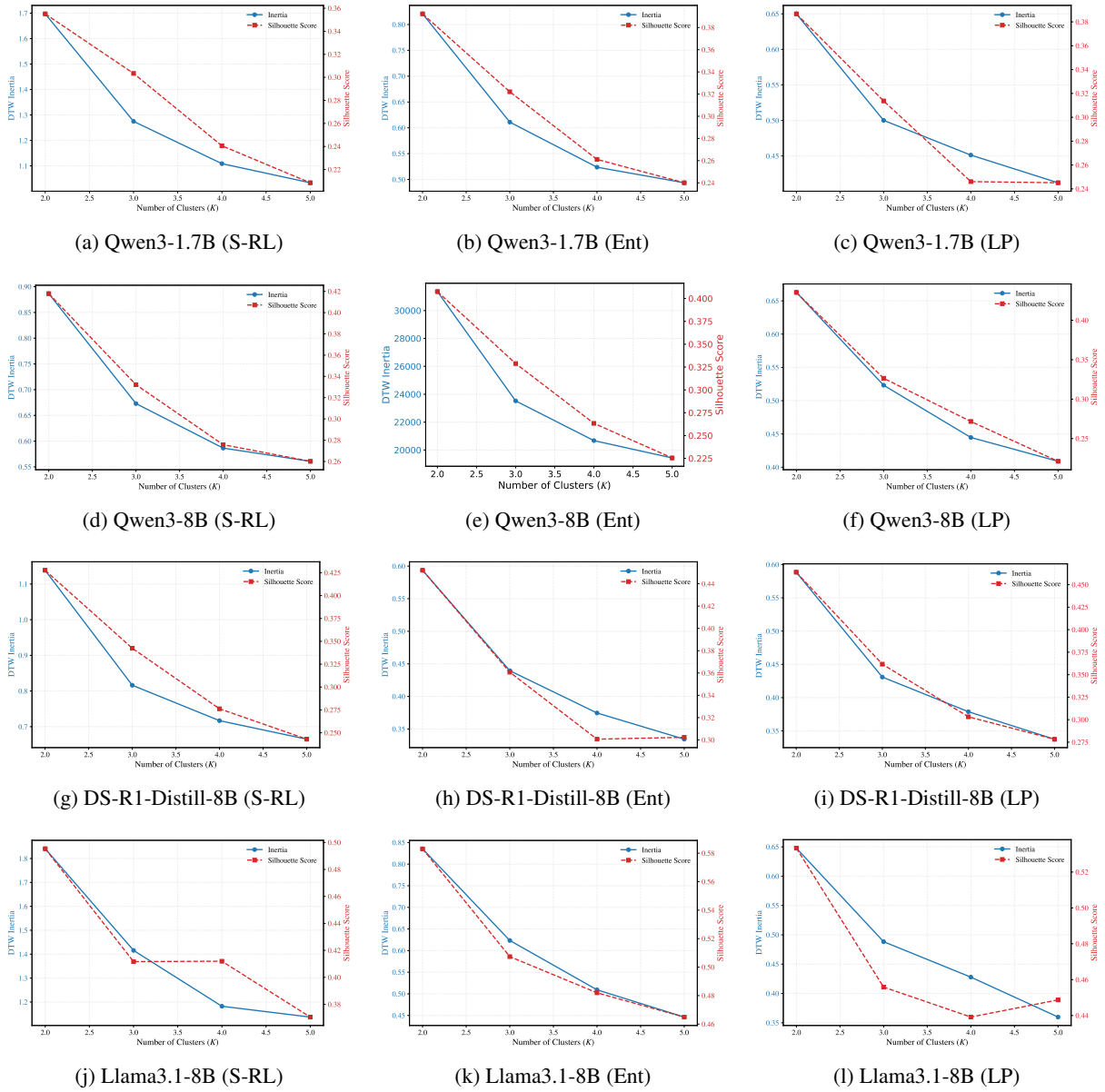
B.4 Token Overlap Analysis Across Entropy-Driven Clusters

As illustrated in Figure 10, we present the overlap bar charts for all four evaluated models. Across different parameter scales and base architectures, the intersection of the top representative tokens among the three entropy-driven clusters remains consistently high.

B.5 Cross-Model Semantic Mapping of Entropy Clusters

As presented in Table 5, these clusters naturally map to consistent semantic profiles, corresponding precisely to low, medium, and high entropy behaviors during mathematical reasoning. The vocabulary stratification unfolds into three universal semantic functions across all architectures:

- **Low-Entropy States (Execution):** Across all configurations, the core neighborhood is



(a) Qwen3-1.7B (S-RL) (b) Qwen3-1.7B (Ent) (c) Qwen3-1.7B (LP)

(d) Qwen3-8B (S-RL) (e) Qwen3-8B (Ent) (f) Qwen3-8B (LP)

(g) DS-R1-Distill-8B (S-RL) (h) DS-R1-Distill-8B (Ent) (i) DS-R1-Distill-8B (LP)

(j) Llama3.1-8B (S-RL) (k) Llama3.1-8B (Ent) (l) Llama3.1-8B (LP)

Figure 8: **Comprehensive K-Validity Evaluation Matrix.** The DTW Inertia and Silhouette Scores are plotted for $K \in [2, 5]$ across all 12 experimental configurations.

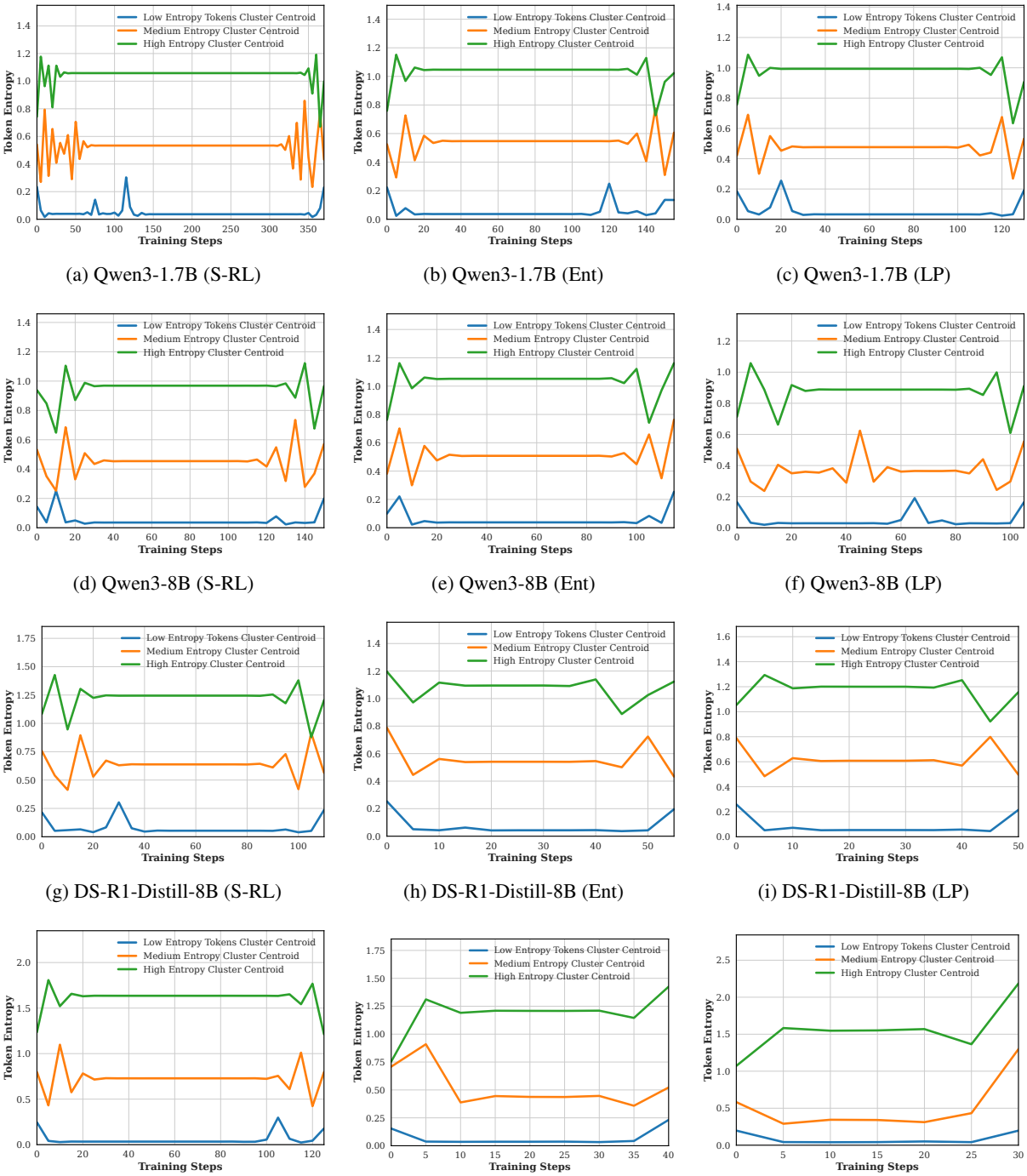
exclusively occupied by concrete arithmetic digits, operators, and formatting syntax (such as LaTeX brackets and structural grammar). This proves that the low-entropy state universally captures deterministic computational execution.

- **Mid-Entropy States (Logic):** The mid-level uncertainty cluster consistently functions as the structural transition backbone, populated by variable establishments (`Let`, `given`) and step-by-step deductive operators (`Therefore`, `Wait`, `Similarly`). It reflects the orderly advancement of a reasoning path.

- **High-Entropy States (Thinking):** The highest entropy states organically gather tokens representing planning, epistemic uncertainty (`perhaps`, `maybe`, `suppose`), and verification triggers (`analyze`, `confirm`, `contrad`). This identifies the model's active state of searching hypotheses or executing self-correction.

Semantic Cluster (Entropy)	Model	Method	Shape A: Iconic Tokens (Closest to Centroid)	Shape B: Core Tokens (Neighborhood Freq.)
Execution (Low) *Math Computation & Formatting*	DeepSeek-R1-Distill-8B	S-RL Ent LP	)/, 16, |, y \, k,]\n, than]\n, c, =, equation	by, 0, 3, 1 \\,]\n, **, ^ of, =, ,, to
	Llama3.1-8B	S-RL Ent LP	b, x, +, 45 $, 25, for, 24)$, *, +, is	1, 6, 3, 4 -, }, $, = of, $, to, is
	Qwen3-1.7B	S-RL Ent LP	+, *, 3, (example, y, k, ^ example, +,)), *	to, =, be,) be, =,), to be,), =, *
	Qwen3-8B	S-RL Ent LP	8, B, b, error ^, \, $, k y, by, (-, than	by,), =, be =, me, \, $$ =, by,), |
Logic (Mid) *Structural Transition & Validation*	DeepSeek-R1-Distill-8B	S-RL Ent LP	Then, that, here, compute again, such, about, correct these, you, says, an	Let, So, but, that ?, here, So, complicated but, 'm, Wait, ?
	Llama3.1-8B	S-RL Ent LP	found, den, follows, some Simpl, at, do, will Sub, add, states, answer	help, need,)., scenario Now, let, In, one has,)., all, at
	Qwen3-1.7B	S-RL Ent LP	don, Yes, let, what But, No, 's, Since Wait, But, ?, I	But, we, there, Wait But, Therefore, So, Wait But, ?, Wait, this
	Qwen3-8B	S-RL Ent LP	Yes, these, another, which Now, check, Similarly).\n\n, about, them, recall	?, but, Wait, it Wait, ?, this, Let Let, but, :\n\n, Then
Thinking (High) *Hypothesis & Divergent Search*	DeepSeek-R1-Distill-8B	S-RL Ent LP	If, Also, since, perhaps wait, Hmm, write, maybe think, Hmm, when, maybe	maybe, let, perhaps, wait maybe, perhaps, wait, But maybe, perhaps, wait, Hmm
	Llama3.1-8B	S-RL Ent LP	Given, First, Calcul, Let now, Let, determine, 're assume, From, So, First	Calcul, Let, When, Thus To, The, will, express First, start, denote, To
	Qwen3-1.7B	S-RL Ent LP	???, probably, under, smarter Thus, confirm, larger, relationship adding, assuming, regardless, What	wait, ???, Divide,)" perhaps, ???, Hence, relationship perhaps, Thus, tackle, Hence
	Qwen3-8B	S-RL Ent LP	Well, Which, setting, lead analyze, contrad, Hmm Thus, express, Sup, Therefore	perhaps, Thus, consider, Hence perhaps, maybe, suppose However, maybe, If, perhaps

Table 5: **Comprehensive Cross-Model Token Semantic Mapping Matrix. Shape A (Iconic)** identifies the single most morphologically standard instance (minimum DTW distance to the centroid) among the top 100 globally frequent tokens. **Shape B (Core)** extracts the most frequently occurring tokens strictly within the closest 10% spatial neighborhood of the centroid.



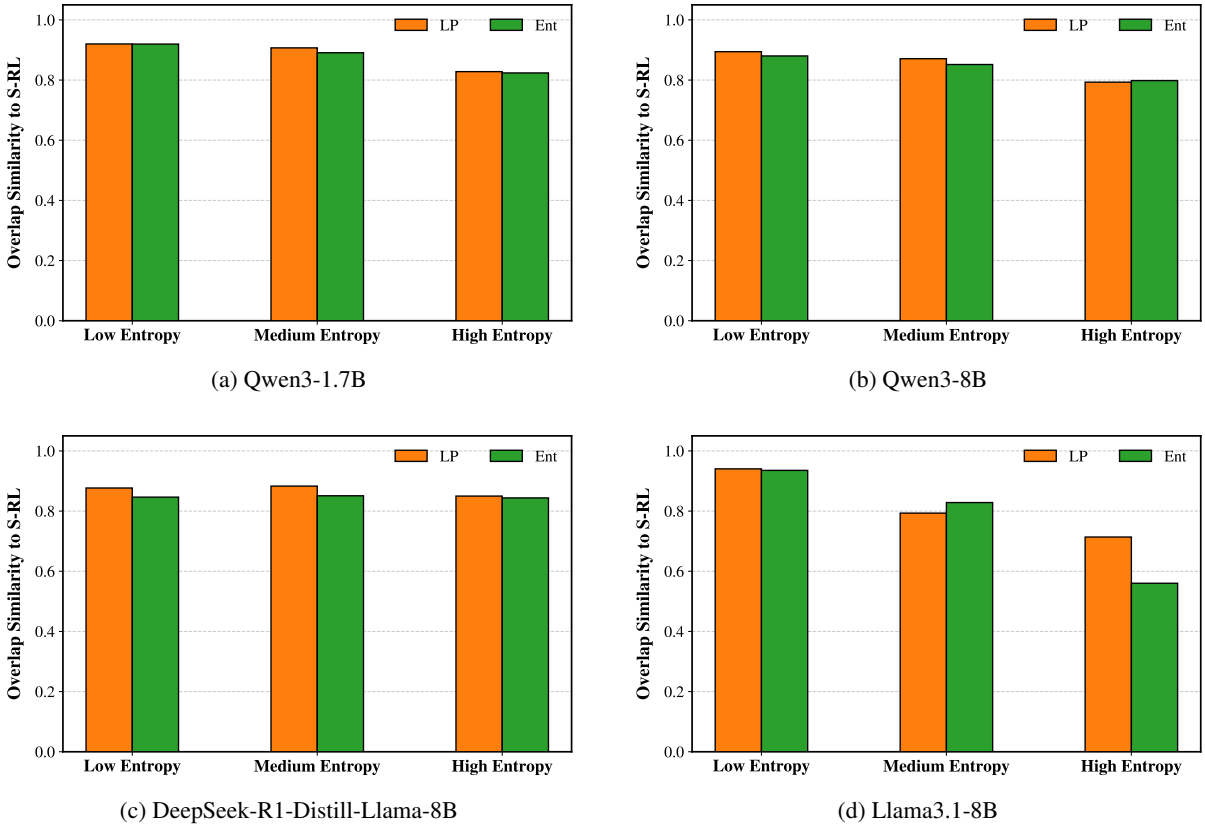
(a) Qwen3-1.7B (S-RL) (b) Qwen3-1.7B (Ent) (c) Qwen3-1.7B (LP)

(d) Qwen3-8B (S-RL) (e) Qwen3-8B (Ent) (f) Qwen3-8B (LP)

(g) DS-R1-Distill-8B (S-RL) (h) DS-R1-Distill-8B (Ent) (i) DS-R1-Distill-8B (LP)

(j) Llama3.1-8B (S-RL) (k) Llama3.1-8B (Ent) (l) Llama3.1-8B (LP)

Figure 9: **Comprehensive Visualization of Entropy-Driven Clusters.** Trajectories of computed DTW centroids are plotted across all configurations.



(a) Qwen3-1.7B (b) Qwen3-8B

(c) DeepSeek-R1-Distill-Llama-8B (d) Llama3.1-8B

Figure 10: **Token Overlap Analysis Across Entropy-Driven Clusters.** The bar charts illustrate the vocabulary overlap among the three identified entropy levels (Low, Mid, and High) for each model.