HEALing Entropy Collapse: Enhancing Exploration in Few-Shot RLVR via Hybrid-Domain Entropy Dynamics Alignment

Zhanyu Liu[1,4*], Qingguo Hu[1,4*†], Ante Wang[2], Chenqing Liu[1,4], Zhishang Xiang[1,4], Hui Li[1,4], Delai Qiu[3], Jinsong Su [1,4‡]

[1]School of Informatics, Xiamen University

[2]Tsinghua University [3]Xiamen Unisound Intelligence Technology Co., Ltd

[4]Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and Taiwan (Xiamen University), Ministry of Culture and Tourism, China

{liuzhanyu, huqingguo}@stu.xmu.edu.cn, jssu@xmu.edu.cn

Abstract

Reinforcement Learning with Verifiable Reward (RLVR) has proven effective for training reasoning-oriented large language models, but existing methods largely assume high-resource settings with abundant training data. In low-resource scenarios, RLVR is prone to more severe entropy collapse, which substantially limits exploration and degrades reasoning performance. To address this issue, we propose **H**ybrid-domain **E**ntropy dynamics **AL**ignment (**HEAL**), a framework tailored for few-shot RLVR. HEAL first selectively incorporates high-value general-domain data to promote more diverse exploration. Then, we introduce Entropy Dynamics Alignment (EDA), a reward mechanism that aligns trajectory-level entropy dynamics between the target and general domains, capturing both entropy magnitude and fine-grained variation. Through this alignment, EDA not only further mitigates entropy collapse but also encourages the policy to acquire more diverse exploration behaviors from the general domain. Experiments across multiple domains show that HEAL consistently improves few-shot RLVR performance. Notably, using only 32 target-domain samples, HEAL matches or even surpasses full-shot RLVR trained with 1K target-domain samples. Our code is available at https://github.com/XMUDeepLIT/HEAL.

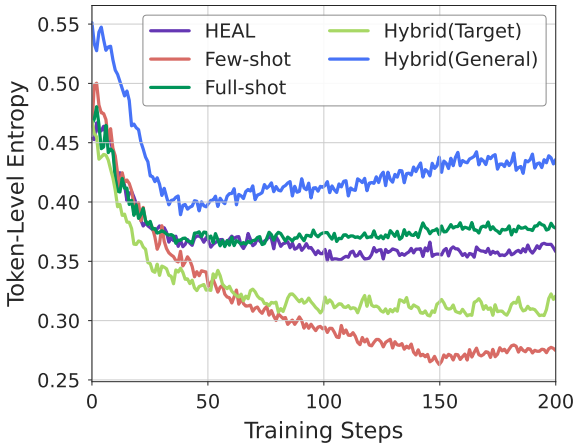

Figure 1: Average token-level entropy during training under different settings. Few-shot and Full-shot denote RLVR trained with few-shot and sufficient target-domain data, respectively. Hybrid(Target) and Hybrid(General) indicate few-shot RLVR augmented with randomly sampled general-domain data, evaluated on target- and general-domain data. Notably, our framework enables few-shot RLVR to achieve entropy magnitude comparable to full-shot training.

1 Introduction

Recent breakthroughs in Large Language Models (LLMs) (Liu et al., 2024a; Yang et al., 2025; Hu et al., 2025a) have given rise to a new generation of reasoning-oriented systems, exemplified by models such as OpenAI o1 (OpenAI, 2024) and DeepSeek-R1 (DeepSeek-AI et al., 2025), which demonstrate substantially improved performance on complex reasoning tasks. A key technique in achieving such success is Reinforcement Learning with Verifiable Reward (RLVR) (Lambert et al., 2024; DeepSeek-AI et al., 2025), which utilizes rule-based outcome rewards to provide a binary feedback on the correctness of a model's final answer. This simple yet effective mechanism not only mitigates reward hacking (Cui et al., 2025a), but also eliminates the need to train complex reward models (Schulman et al., 2017).

Despite its demonstrated effectiveness, existing RLVR research (Shao et al., 2024; Yu et al., 2025; Yue et al., 2025b; Liu et al., 2025a) has predominantly focused on high-resource domains, where large volumes of high-quality training samples are readily available. However, this assumption of abundant and reliable reward signals does not hold in many real-world domains, such as medical

*Equal contribution.
†Project lead.
‡Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30732–30748
July 2-7, 2026

reasoning (Zhang et al., 2025a) and other specialized knowledge domains (Guha et al., 2023; Zhang et al., 2025b), where training data for RLVR are often scarce. In this context, the policy is prone to rapid overfitting to a few generated trajectories, which prematurely limits exploration and leads to more severe entropy collapse compared to high-resource scenarios (Yue et al., 2025a; Cui et al., 2025b). As shown in Figure 1, compared to sufficient training data (Full-shot), Few-shot RLVR exhibits a significantly lower entropy magnitude. While recent studies (He et al., 2025; Zheng et al., 2025) have proposed various methods to mitigate entropy collapse in RLVR, they often overlook the scarcity of training data, and directly applying them under such conditions may be sub-optimal.

In this paper, we propose **H**ybrid-domain **E**ntropy dynamics **AL**ignment (**HEAL**), a novel framework specifically designed to boost exploration diversity for RLVR under low-resource scenarios. Our proposed framework is built upon two key components:

We firstly incorporate readily available general-domain data. Intuitively, although general-domain data may not provide domain-specific knowledge, it offers fundamental reasoning patterns and thus encourages more diverse exploration. Much like a human learner applying general skills to a new domain (Gick and Holyoak, 1980), this hybrid training prevents the policy from prematurely narrowing its search space, thereby notably mitigating entropy collapse in the target domain. To further avoid incurring excessive training costs from general-domain data, we adopt a data selection strategy that retains only a small set of high-value samples based on their reasoning uncertainty and exploratory diversity.

Despite the benefits of this hybrid training, we find that the policy's entropy in the target domain remains substantially lower than that of the general domain. To bridge this gap, we then introduce a novel reward mechanism, termed Entropy Dynamics Alignment (EDA). Unlike conventional approaches that naively increase entropy without constraints (Wang et al., 2025b; Liu et al., 2025b), which can limit policy exploitation or even destabilize training, EDA leverages general-domain data as a reference to guide the policy toward more diverse exploration in target domain. Specifically, EDA constructs *trajectory-level entropy dynamics* for both target- and general-domain data, which captures not only the token-level entropy magnitude but also fine-grained variation. By comparing entropy dynamics within and across domains, EDA rewards trajectories that exhibit stronger inter-domain similarity, thereby effectively encouraging alignment between target- and general-domain entropy characteristics (magnitude and variation). Through this mechanism, the policy not only elevates entropy in the target domain in a controlled manner but also acquires more diverse exploratory behaviors guided by the general domain.

To demonstrate the effectiveness of our proposed framework, we conduct extensive experiments across multiple domains, including Medicine, Physics, Code, and Math. Experiments on the Qwen3 (Yang et al., 2025) and LLaMA-3.2 (Dubey et al., 2024) series of models show that our framework consistently and substantially improves the performance of few-shot RLVR. Remarkably, with only 32 target-domain samples, our approach matches or even surpasses full-shot RLVR trained with 1K target-domain samples. Further analyses indicate that our framework also outperforms existing entropy regularization methods in low-resource scenarios.

2 Related Work

Low-Resource RLVR Extensive research has explored Supervised Fine-Tuning (SFT) (Chen et al., 2023; Ivison et al., 2025) and Reinforcement Learning from Human Feedback (RLHF) (Muldrew et al., 2024; Liu et al., 2024b; Das et al., 2025) as primary paradigms for improving the performance of LLMs under low-resource settings. Despite their success, how to effectively apply RLVR in such scenarios remains an open problem. To address this challenge, emergent paradigms leverage self-play and autonomous feedback (Huang et al., 2025; Zhao et al., 2025; Hu et al., 2025b) to enhance reasoning capabilities without human annotations. However, due to the lack of reliable supervision signals, such self-improvement is often constrained by the model's internal knowledge boundaries, thereby limiting out-of-domain generalization. Recent data-centric studies (Li et al., 2025) have demonstrated the potential of leveraging a very small amount of data to boost the reasoning ability of LLMs (i.e., few-shot RLVR). Nevertheless, exploration of this paradigm remains preliminary, and its associated challenges have yet to be fully addressed. For instance, Wang et al. (2025b) reports severe training collapse when ap-

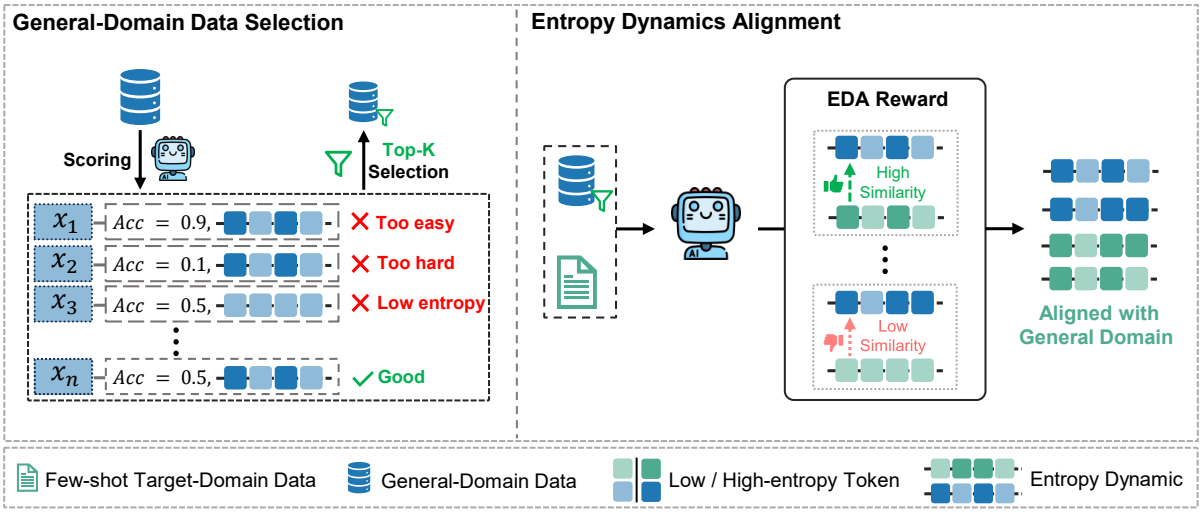

Figure 2: Overview of our proposed HEAL framework. *Left*: We incorporate a small set of high-value general-domain data into few-shot RLVR to promote diverse exploration and mitigate entropy collapse in the target domain. *Right*: Entropy Dynamics Alignment reward guides the policy by aligning the trajectory-level entropy dynamics of few-shot target-domain data with those of the selected general-domain data, thereby further encouraging controlled increases in entropy and more diverse exploratory behaviors.

plying 1-shot RLVR under certain settings, without identifying the underlying mechanisms. To bridge this gap, our work reveals the challenges of entropy collapse in few-shot RLVR and proposes effective solutions to mitigate this issue.

Entropy Perspectives in RLVR Recent RLVR research (Yue et al., 2025a; Dang et al., 2025; Min et al., 2026) highlights the role of entropy in the exploration–exploitation trade-off, showing that higher entropy promotes diverse reasoning trajectories, whereas lower entropy facilitates convergence. However, in RLVR, entropy often exhibits excessive contraction, causing the policy to lose exploratory capacity, which is referred to as entropy collapse (Cui et al., 2025b; He et al., 2025; Liu et al., 2025b; Wu et al., 2025). To mitigate entropy collapse during RLVR, existing studies (Sheng et al., 2024; He et al., 2025; Wang et al., 2025a; Cheng et al., 2025; Chen et al., 2025b) have explored a range of regularization-based strategies to preserve exploratory diversity. Nevertheless, these works typically do not account for the effect of data scale. In contrast, we show that entropy collapse becomes substantially more severe under low-resource scenarios. To address this issue, our proposed HEAL introduces hybrid-domain training and guides the policy to align fine-grained entropy characteristics across domains, thereby substantially mitigating entropy collapse and promoting diverse exploration.

3 Our Proposed Framework

In this section, we present **H**ybrid-domain **E**ntropy dynamics **AL**ignment (**HEAL**), a novel framework designed to mitigate entropy collapse in RLVR under low-resource scenarios. As illustrated in Figure 2, HEAL consists of two key components: We first incorporate carefully selected general-domain data into the training process. By doing so, we can effectively introduce more diverse exploratory behaviors beyond the target domain, preventing the policy from prematurely collapsing its search space (§3.1). To further promote exploration in the target domain, we then introduce Entropy Dynamics Alignment (EDA), which guides the policy to align the entropy characteristics between the target and general domains (§3.2).

3.1 Hybrid Training with Selected General-Domain Data

When encountering a new domain, human learners often demonstrate remarkable transfer learning capabilities by leveraging general skills and problem-solving strategies (Gick and Holyoak, 1980), even with limited prior knowledge. Building on this insight, we incorporate readily available general-domain data (e.g., commonsense reasoning) into the few-shot RLVR training process. Although these samples lack domain-specific knowledge, they provide fundamental reasoning patterns and thus promote more diverse exploration.

To validate this intuition, we randomly selected 1K samples from a commonsense-reasoning dataset (Talmor et al., 2019) and directly combined them with few-shot target-domain data for RLVR training. As illustrated in Figure 1, this simple mixture notably alleviates entropy collapse in the target domain and leads to notable performance improvements compared to training only on few-shot target-domain data.

However, indiscriminately incorporating excessive general-domain data would be prohibitively expensive in terms of computational cost. Therefore, to maximize the efficacy of this hybrid training, we select only a small set of high-value general-domain samples based on two complementary criteria: *Reasoning Uncertainty* and *Exploratory Diversity*.

- *Reasoning Uncertainty*: Given N generated trajectories $\{y_1, \ldots, y_N\}$ for an input question x, we first compute the average accuracy of these trajectories as $\mathrm{Acc}(x) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(y_i = y_{\text{gt}})$, where $\mathbb{I}(\cdot)$ denotes the indicator function and y_{gt} is the ground-truth answer. We then define the reasoning uncertainty as $\mathrm{Uncertainty}(x) = 1 - 2\left|\mathrm{Acc}(x) - \frac{1}{2}\right|$. Samples with high uncertainty are preferred, as they have been shown to be the most effective for enhancing reasoning capabilities (Huang et al., 2025).

- *Exploratory Diversity*: To select samples that elicit more diverse exploration, we collect the top 20% of tokens with the highest entropy from each generated trajectory and compute their average entropy as the exploratory diversity, denoted as $\mathrm{Diversity}(x)$. A higher exploratory diversity score indicates that the sample induces more varied exploratory behaviors (Wang et al., 2025a).

Finally, we combine the two criteria into a single scalar score:

$$c(x) = \mathrm{Uncertainty}(x) \cdot \mathrm{Diversity}(x). \quad (1)$$

We select the top-K general-domain samples with the highest composite scores to serve as high-quality training data.

3.2 Entropy Dynamics Alignment

While hybrid training helps reduce entropy collapse in the target domain, we observe that the policy's entropy in the target domain remains considerably lower than that in the general domain, as shown in Figure 1. This gap indicates that merely mixing general-domain data is insufficient to mitigate entropy collapse in the target domain. To bridge this gap, we propose Entropy Dynamics Alignment (EDA), a novel reward mechanism designed to guide the policy toward more diverse exploration in the target domain by aligning its entropy characteristics with those observed in the general domain.

Trajectory-Level Entropy Dynamics We first provide the definition of trajectory-level entropy dynamics. Formally, given a trajectory y_i, we define its entropy dynamics as the sequence of token-level entropies over generation steps, $\tau_y = (\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_{|y_i|})$, where H_t is the entropy over the vocabulary distribution at timestep t, conditioned on the input and previously generated tokens. This sequence captures not only the token-level entropy magnitude of a trajectory, but also its fine-grained variation, thereby providing a richer representation than direct aggregation (He et al., 2025; Cheng et al., 2025).

The discrepancy between the entropy dynamics of two trajectories can be measured using a suitable similarity function $s(\cdot, \cdot)$. Specifically, given two trajectories y_1 and y_2, the discrepancy between their entropy dynamics is defined as $s(\tau_{y_1}, \tau_{y_2})$. Since trajectories may vary in length, we first apply an interpolation-based alignment strategy to normalize their lengths before computing the similarity. Implementation details are provided in the Appendix A.2.

Entropy Dynamics Alignment Reward The core idea of this reward is to align the entropy characteristics between target-domain and general-domain data, in terms of both token-level magnitude and fine-grained variation. During hybrid training, we collect entropy dynamics from both the target and general domains, denoted as $\mathcal{B}_{\text{tgt}}$ and $\mathcal{B}_{\text{gen}}$, respectively. For a target-domain entropy dynamics $\tau_{y_i} \in \mathcal{B}_{\text{tgt}}$, we compute two kinds of similarity measures.

- *Intra-Domain Similarity*: We define the intra-domain similarity as the maximum similarity between τ_i and other entropy dynamics from the same domain:

$$\mathcal{S}_{\text{intra}}(\tau_{y_i}) = \max_{\tau_{y_j} \in \mathcal{B}_{\text{tgt}},\, j \neq i} s(\tau_{y_i}, \tau_{y_j}). \quad (2)$$

- *Inter-Domain Similarity*: Similarly, we define the inter-domain similarity as the maximum similarity between τ_i and entropy dynamics from the general domain:

$$\mathcal{S}_{\text{inter}}(\tau_{y_i}) = \max_{\tau_{y_k} \in \mathcal{B}_{\text{gen}}} s(\tau_{y_i}, \tau_{y_k}). \quad (3)$$

We then assign a reward to trajectories whose entropy dynamics exhibit higher inter-domain similarity than intra-domain similarity. These trajectories are considered strong exemplars for promoting alignment across domains. Based on this, we define the EDA reward for the trajectory y_i as a binary signal:

$$r_{\text{EDA}}(y_i) = \begin{cases} 1, & \text{if } \mathcal{S}_{\text{inter}}(\tau_{y_i}) > \mathcal{S}_{\text{intra}}(\tau_{y_i}), \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

By doing so, this reward mechanism not only elevates entropy in the target domain in a controlled manner but also promotes fine-grained alignment of entropy variation, allowing the policy to implicitly acquire more diverse exploratory behaviors from the general domain.

Policy Update Standard RLVR uses a deterministic accuracy reward, denoted as $r_{\text{Acc}}(y_i)$, which assesses whether the final answer of trajectory y_i matches its ground truth. In our framework, we combine this accuracy reward with our proposed EDA reward. Formally, the final reward for trajectory y_i is defined as $r(y_i) = r_{\text{Acc}}(y_i) + r_{\text{EDA}}(y_i)$. Consequently, the policy optimization objective is formulated as minimizing the following loss function:

$$\mathcal{L}_{\text{RLVR}}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(\cdot|x)} [r(y)] \quad (5)$$

where θ represents the policy parameters to be optimized, $\mathcal{D}$ denotes the training dataset containing prompts x, $\pi_\theta(\cdot|x)$ is the likelihood of the generated trajectory y.

4 Experiments

4.1 Setup

Datasets We evaluate our framework across three domains: *Medicine*, *Physics*, and *Code*. These domains have relatively scarce open-source and high-quality data, particularly in Medicine and Physics. For the Medicine domain, we utilize MedBullets (Chen et al., 2025a) and MedXpertQA (Zuo et al., 2025), following the data-splitting pipeline proposed by Qiu et al. (2025) to construct the training and test sets. For the Physics domain, we use WebInstruct (Ma et al., 2025) for training. For the Code domain, we employ LiveCodeBench (v1–v4) (Jain et al., 2024) as the training dataset. To simulate real-world RLVR under low-resource scenarios, we randomly sample a small number of training samples from each domain to construct few-shot datasets. This setup enables an effective evaluation of our framework under few-shot settings and facilitates direct comparison with the high-resource scenarios.

For the general-domain dataset, we adopt CommonsenseQA (Talmor et al., 2019) for hybrid training. CommonsenseQA is a multiple-choice benchmark designed to evaluate models' commonsense reasoning over everyday scenarios, encompassing background world knowledge, causal relations, and social reasoning. Therefore, it exhibits minimal overlap with the knowledge required in the aforementioned target domains.

Models and Training Details To evaluate the generalizability of our framework, we perform RLVR across multiple models with different architectures, including Qwen3-1.7B-Base, Qwen3-4B-Base (Yang et al., 2025), and LLaMA-3.2-3B-Instruct (Dubey et al., 2024). We adopt VERL (Sheng et al., 2024) as the RLVR training pipeline. For all models, we sample 8 trajectories per input. The rollout temperature is set to 0.7 for Qwen3 and 0.6 for LLaMA-3.2, following Wang et al. (2025b). The training batch size and mini-batch size are both set to 128, with a maximum input length of 1024 tokens and a maximum trajectory length of 3072 tokens. For the remaining hyperparameters, we follow the training recipe of GRPO (Shao et al., 2024). All experiments are conducted on 8 × A100 (80 GB) GPUs.

Evaluation Benchmarks For the Medicine domain, evaluation is conducted on the test splits of MedBullets (Mbul.) and MedXpertQA (MedXQA). For the Physics domain, we evaluate on the physics subsets of both C-Eval (Huang et al., 2023) and WebInstruct (WebIns.) (Ma et al., 2025). For both the Medicine and Physics domains, we adopt the Qwen2.5-Math (Yang et al., 2024) evaluation pipeline and report the Avg@4 score. For the Code domain, we evaluate on LiveCodeBench v5 (LCBv5) (Jain et al., 2024) and HumanEval Plus (HEval+) (Chen et al., 2021), using the official LiveCodeBench (Jain et al., 2024)

	Data Size		Medicine			Physics			Code		
	$\lvert\mathcal{D}_{tgt}\rvert$	$\lvert\mathcal{D}_{gen}\rvert$	MBul.	MedXQA	Avg.	C-Eval	WebIns.	Avg.	HEval+	LCBv5	Avg.
Qwen3-1.7B-Base			41.78	31.06	36.42	49.53	5.69	27.61	55.49	5.39	30.44
↪ Full-shot	1K	0	49.01	**40.45**	**44.73**	60.68	11.79	36.24	60.98	17.37	39.18
↪ Few-shot	32	0	45.07	38.33	41.70	57.27	7.32	32.30	56.71	12.57	34.64
↪ Only-General	0	10K	42.11	32.73	37.42	55.39	7.72	31.56	58.54	10.18	34.36
↪ Hybrid	32	384	45.39	38.64	42.02	59.74	8.13	33.94	57.93	13.77	35.85
↪ **HEAL**	32	384	**49.34**	40.00	44.67	**62.19**	**12.20**	**37.20**	**62.80**	**19.76**	**41.28**
Qwen3-4B-Base			43.09	33.93	38.51	53.88	8.94	31.41	66.46	16.77	41.62
↪ Full-shot	1K	0	**68.09**	**45.30**	**56.70**	74.10	**15.04**	44.57	78.05	23.95	51.00
↪ Few-shot	32	0	55.26	43.18	49.22	73.35	11.79	42.57	71.95	22.75	47.35
↪ Only-General	0	10K	49.67	39.70	44.69	68.62	8.13	38.38	70.12	20.35	45.24
↪ Hybrid	32	384	55.92	42.73	49.33	76.75	11.38	44.07	74.39	24.55	49.47
↪ **HEAL**	32	384	57.24	44.39	50.82	**80.15**	13.82	**46.99**	**79.88**	**26.95**	**53.42**
LLaMA3.2-3B-Instruct			40.46	26.82	33.64	33.46	4.88	19.17	48.17	11.38	29.78
↪ Full-shot	1K	0	**84.87**	**42.88**	**63.88**	36.67	6.91	21.79	51.22	16.17	33.70
↪ Few-shot	32	0	56.91	31.36	44.14	35.54	6.10	20.82	50.61	12.57	31.59
↪ Only-General	0	10K	48.68	28.48	38.58	36.11	5.69	20.90	51.22	13.77	32.50
↪ Hybrid	32	384	58.55	30.15	44.35	35.16	5.28	20.22	51.83	15.57	33.70
↪ **HEAL**	32	384	60.53	32.27	46.40	**37.05**	**7.72**	**22.39**	**52.44**	**16.77**	**34.61**

Table 1: Comprehensive results on reasoning benchmarks across three target-domains. $|\mathcal{D}_{tgt}|$ and $|\mathcal{D}_{gen}|$ denote the sizes of the target- and general-domain training data, respectively. The best results are highlighted in bold, and the second-best results are underlined.

evaluation pipeline and EvalPlus (Liu et al., 2023). We report Pass@10 and Pass@1 scores for LiveCodeBench v5 and HumanEval Plus, respectively.

Baselines To better evaluate the effectiveness of our framework, we compare HEAL against the following baselines: (1) Full-shot: trained on sufficient target-domain data sampled from each domain's training set, serving as a high-resource performance upper bound; (2) Few-shot: trained only on the few-shot target-domain samples; (3) Only-General: trained only on large-scale general-domain samples; (4) Hybrid: naively combines the target-domain samples with randomly selected general-domain samples.

4.2 Main Results

Table 1 presents the comprehensive performance of HEAL compared to various baselines across three target domains. Based on the overall evaluation results, we highlight several key conclusions:

HEAL Significantly Improves the Performance of Few-Shot RLVR Our empirical results demonstrate that the HEAL framework yields substantial performance gains over the vanilla Few-shot baseline across all evaluated models and benchmarks. For instance, HEAL improves the average scores of Qwen3-1.7B-Base over the Few-shot baseline by up to 6.64% in the Code domain. Notably, our framework consistently outperforms the Hybrid baseline, demonstrating the effectiveness of general-domain data selection and EDA. From the perspective of entropy, HEAL effectively mitigates entropy collapse in few-shot RLVR, enabling the policy to achieve an entropy magnitude comparable to that of full-shot training, as shown in Figure 1.

HEAL Enables Few-Shot RLVR to Match or Even Surpass Full-Shot Performance Remarkably, using only 32 target-domain samples, HEAL matches or even surpasses the performance of models trained on 1K samples (Full-shot), with this effect particularly pronounced in the Physics and Code domains. For example, Qwen3-4B-Base with HEAL achieves average scores of 46.99% in Physics and 53.42% in Code, outperforming its Full-shot counterpart (44.57% and 51.00%, respectively). A similar trend is also evident for LLaMA3.2-3B-Instruct, which surpasses its Full-shot performance in both domains. These results further underscore HEAL's advancement in low-resource scenarios, demonstrating its ability to achieve competitive performance with minimal target-domain data.

The Role of General-Domain Data The results of the Only-General baseline reveal that even leveraging massive general-domain samples often yields limited performance gains, consistently underperforming the Few-shot baseline (e.g., Avg.

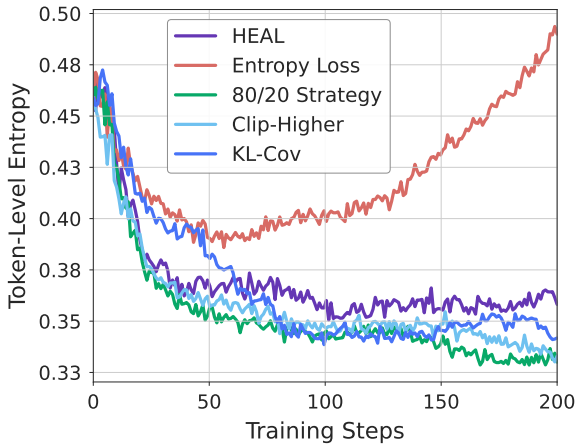

Figure 3: Average token-level entropy during training under different entropy regularization methods and HEAL. We conduct experiments on the Qwen3-1.7B-Base model using the same training dataset.

Method	Medicine	Physics	Code
Entropy Loss	42.79	34.43	38.27
80/20 Strategy	42.47	<u>36.36</u>	<u>40.08</u>
Clip-Higher	42.59	34.52	39.18
KL-Cov	<u>44.00</u>	35.58	38.57
HEAL	**44.67**	**37.20**	**41.28**

Table 2: Performance comparison between HEAL and other entropy regularization methods. We report the average score for each domain.

44.69% vs. 49.22% in Medicine for Qwen3-4B-Base). This demonstrates that the introduced general-domain data does not directly provide domain-specific knowledge. Instead, it serves to supply more fundamental reasoning patterns. These results further highlight that the performance gains achieved by HEAL arise from its effective utilization of the diverse exploratory behaviors present in general-domain data, rather than from directly injecting domain knowledge to artificially boost performance.

4.3 Analysis

Comparison with Existing Entropy Regularization Methods Many recent studies have proposed various entropy regularization methods to mitigate entropy collapse in RLVR. However, these methods typically do not account for data scarcity. To rigorously evaluate our framework, we compare HEAL with four representative entropy regularization baselines: *Entropy Loss* (He et al., 2025), *80/20 Strategy* (Wang et al., 2025a), *Clip-Higher* (Yu et al., 2025), and *KL-Cov* (Cui et al., 2025b). Briefly, Entropy Loss directly encourages higher policy entropy through a dedicated loss term, 80/20 Strategy selects the highest-entropy tokens for policy updates, Clip-Higher increases the likelihood of low-probability exploration tokens, and KL-Cov applys KL penalty to tokens with high covariances. Details about these methods are provided in Appendix A.3.

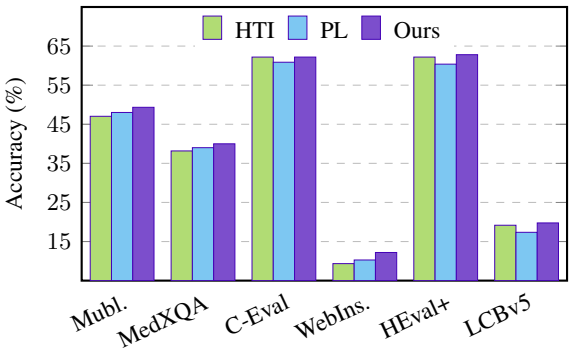

Figure 4: Performance comparison using different similarity functions. Experiments are conducted on the Qwen3-1.7B-Base model.

As shown in Table 2, our framework consistently outperforms all baselines across the three target domains. This superior performance can be attributed to HEAL's more effective mitigation of policy entropy collapse. As illustrated in Figure 3, the standard Entropy Loss, which naively increases entropy without constraints, can even lead to entropy explosion in later stages of training. For the 80/20 Strategy, Clip-Higher, and KL-Cov, their ability to mitigate entropy collapse is limited in low-resource scenarios. In contrast, HEAL leverages general-domain data as a reference to guide the policy, enabling more controlled adjustment of entropy magnitude while also utilizing learned fine-grained variation to effectively promote more diverse exploratory behaviors.

Impact of Different Similarity Measures for Entropy Dynamics In addition to the KL divergence used in our final implementation, we also experiment with two alternative strategies: High-entropy Tokens Intersection (HTI) and Pearson Linear (PL) similarity. Briefly, HTI measures the overlap between high-entropy tokens, while PL computes the correlation of slopes after linearly fitting the trajectory-level entropy dynamics. Implementation details are provided in Appendix A.2. As shown in Figure 4, the KL divergence used in our framework consistently outperforms the other two methods, demonstrating its superior ability to capture fine-grained differences in trajectory-level

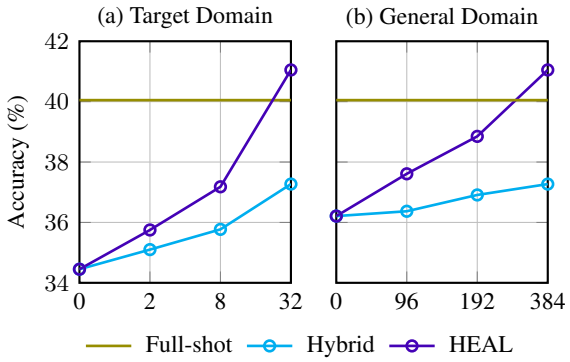

Figure 5: Average accuracy of the Qwen3-1.7B-Base model across three target domains with respect to different data sizes: (a) varying the target-domain data size while keeping the general-domain data fixed at 384 samples; (b) varying the general-domain data size while keeping the target-domain data fixed at 32 samples.

entropy dynamics. This precise measurement enables more effective alignment, allowing the policy to acquire more diverse exploratory behaviors from the general domain.

Effect of Training Data Scale Figure 5 illustrates the impact of training data scale in both the target and general domains on the performance of the Qwen3-1.7B-Base model. In the target-domain experiment, HEAL surpasses the Hybrid baseline at every data size and approaches or even exceeds the Full-shot upper bound with as few as 32 samples, highlighting its ability to effectively leverage scarce samples. In the general-domain experiment, the improvement of HEAL is also substantially stronger than that of the Hybrid baseline, indicating that it can more effectively transfer fundamental reasoning patterns and diverse exploratory behaviors from general-domain data to the target domain.

Adaptation to Math Domain Although HEAL is specifically designed for low-resource domains, it inherently offers a data-efficient solution under limited computational resources. Therefore, we extend our framework to a high-resource domain (i.e., Math) to assess the efficiency of HEAL. As shown in Table 3, HEAL still demonstrates consistently strong performance. Notably, compared with Full-shot training, HEAL matches or even surpasses its performance while using only 40% of the training data. For instance, on AMC23 and Minerva, HEAL exceeds the Full-shot baseline by 3.49% and 2.57%, respectively. These results underscore HEAL's versatility as a data-efficient RLVR approach for high-resource domains.

	AMC23	Math500	Minerva	Olym.
Qwen3-1.7B-Base	30.94	55.6	13.24	21.48
↪ Full-shot	35.31	**66.2**	19.12	**29.48**
↪ Few-shot	34.06	58.4	17.28	23.70
↪ Only-General	29.38	62.8	16.91	26.07
↪ Hybrid	32.50	60.2	17.65	24.15
↪ HEAL	**38.80**	64.4	**21.69**	25.78

Table 3: Performance of Qwen3-1.7B-Base across four Math domain benchmarks. Details of datasets are in Appendix B, and additional results are in Appendix C.2.

Data Selection		EDA	Med.	Phy.	Code
Uncertainty	Diversity				
×	×	×	42.02	33.94	35.85
✓	×	×	41.49	34.28	36.45
×	✓	×	42.16	34.53	37.04
×	×	✓	43.80	36.44	39.47
✓	✓	×	43.04	35.34	37.36
✓	✓	✓	**44.67**	**37.20**	**41.28**

Table 4: Ablation study of the Qwen3-1.7B-Base model across the Medicine (Med.), Physics (Phy.), and Code domains. Average scores are reported.

4.4 Ablation Study

The ablation study of our framework is summarized in Table 4. We systematically evaluate the contributions of the proposed general-domain data selection strategy and Entropy Dynamics Alignment (EDA). When none of the components is applied, the model exhibits limited performance across all three target domains. Introducing either uncertainty-based or diversity-based data selection leads to only marginal improvements, while combining the two yields more consistent gains, suggesting that these criteria are complementary in identifying high-value general-domain samples. Notably, enabling EDA alone results in a substantial performance boost, underscoring its critical role in mitigating entropy collapse in few-shot RLVR. Overall, the results demonstrate that both high-quality general-domain data selection and EDA are essential to the effectiveness of HEAL, and that their combination produces clear synergistic gains under low-resource scenarios.

5 Conclusion

In this work, we proposed HEAL, a novel framework for mitigating entropy collapse in few-shot RLVR. By integrating selected general-domain data and introducing Entropy Dynamics Alignment, HEAL effectively encourages more diverse policy behaviors. Extensive experiments across

multiple domains demonstrate that HEAL consistently improves few-shot RLVR performance, even matching or surpassing full-shot training with significantly fewer samples. These results highlight HEAL's effectiveness as a data-efficient approach for RLVR in low-resource scenarios.

Limitations

Despite our results are promising, several limitations remain. First, due to computational constraints, HEAL has not been evaluated on larger models, leaving its scalability to be validated. Second, incorporating general-domain data introduces some overhead, but it is limited since only a small amount is used, and we consider it acceptable, as our framework can match or even surpass performance using much more target-domain data.

Acknowledgements

The project was supported by National Key R&D Program of China (No. 2022ZD0160501), Natural Science Foundation of Fujian Province of China (No. 2024J011001), and the Public Technology Service Platform Project of Xiamen (No.3502Z20231043). We also thank the reviewers for their insightful comments.

References

Art of Problem Solving. Amc problems and solutions. https://artofproblemsolving.com/wiki/index.php?title=AMC_Problems_and_Solutions. Accessed: 2025-04-20.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025a. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and 1 others. 2023. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. 2025b. Seed-grpo: Semantic entropy enhanced grpo for uncertainty-aware policy optimization. *arXiv preprint arXiv:2505.12346*.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. 2025. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, and 1 others. 2025a. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, and 1 others. 2025b. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*.

Xingyu Dang, Christina Baek, Kaiyue Wen, Zico Kolter, and Aditi Raghunathan. 2025. Weight ensembling improves reasoning in language models. *arXiv preprint arXiv:2504.10478*.

Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. 2025. Active preference optimization for sample efficient active prefer- 677 ence optimization for sample efficient rlhf. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 96–112.

Daya Guo DeepSeek-AI, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Mary L. Gick and Keith J. Holyoak. 1980. Analogical problem solving. *Cognitive Psychology*, 12(3):306–355.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal

scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, and 1 others. 2025. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*.

Qingguo Hu, Zhenghao Lin, Ziyue Yang, Yucheng Ding, Xiao Liu, Yuting Jiang, Ruizhe Wang, Tianyu Chen, Zhongxin Guo, Yifan Xiong, and 1 others. 2025a. Sigma-moe-tiny technical report. *arXiv preprint arXiv:2512.16248*.

Qingguo Hu, Ante Wang, Jia Song, Delai Qiu, Qingsong Liu, and Jinsong Su. 2025b. Boosting visual knowledge-intensive training for lvlms through causality-driven visual object completion. *arXiv preprint arXiv:2508.04453*.

Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. 2025. R-zero: Self-evolving reasoning llm from zero data. *arXiv preprint arXiv:2508.05004*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*.

Hamish Ivison, Muru Zhang, Faeze Brahman, Pang Wei Koh, and Pradeep Dasigi. 2025. Large-scale data selection for instruction tuning. *arXiv preprint arXiv:2503.01807*.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, and 1 others. 2024. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and 1 others. 2022. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*.

Xuefeng Li, Haoyang Zou, and Pengfei Liu. 2025. Limr: Less is more for rl scaling. *arXiv preprint arXiv:2502.11886*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Bingshuai Liu, Ante Wang, Zijun Min, Liang Yao, Haibo Zhang, Yang Liu, Xu Han, Peng Li, Anxiang Zeng, and Jinsong Su. 2025a. Spec-rl: Accelerating on-policy reinforcement learning with speculative rollouts. *arXiv preprint arXiv:2509.23232*.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. Is your code generated by chatGPT really correct? rigorous evaluation of large language models for code generation. *Advances in Neural Information Processing Systems*.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. 2025b. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models. *arXiv preprint arXiv:2505.24864*.

Zijun Liu, Boqun Kou, Peng Li, Ming Yan, Ji Zhang, Fei Huang, and Yang Liu. 2024b. Enabling weak llms to judge response reliability via meta ranking. *arXiv preprint arXiv:2402.12146*.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Li Erran Li, and 1 others. 2025. Deepscaler: Surpassing o1-preview with a 1.5 b model by scaling rl. *Notion Blog*.

Xueguang Ma, Qian Liu, Dongfu Jiang, Ge Zhang, Zejun MA, and Wenhu Chen. 2025. General-Reasoner: Advancing LLM reasoning across all domains. *Advances in Neural Information Processing Systems*.

Zijun Min, Bingshuai Liu, Ante Wang, Long Zhang, Anxiang Zeng, Haibo Zhang, and Jinsong Su. 2026. Orchestrating tokens and sequences: Dynamic hybrid policy optimization for rlvr. In *Findings of the Association for Computational Linguistics: ACL 2026*.

William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. 2024. Active preference learning for large language models. *arXiv preprint arXiv:2402.08114*.

OpenAI. 2024. Learning to reason with LLMs. Accessed: 2025-04-10.

Zhongxi Qiu, Zhang Zhang, Yan Hu, Heng Li, and Jiang Liu. 2025. Open-medical-r1: How to choose data for rlvr training at medicine domain. *arXiv preprint arXiv:2504.13950*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv 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. *arXiv preprint arXiv:2402.03300*.

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

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Vivek Verma, David Huang, William Chen, Dan Klein, and Nicholas Tomlin. 2025. Measuring general intelligence with generated games. *arXiv preprint arXiv:2505.07215*.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, and 1 others. 2025a. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*.

Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, and 1 others. 2025b. Reinforcement learning for reasoning in large language models with one training example. *arXiv preprint arXiv:2504.20571*.

Junkang Wu, Kexin Huang, Jiancan Wu, An Zhang, Xiang Wang, and Xiangnan He. 2025. Quantile advantage estimation for entropy-safe reasoning. *arXiv preprint arXiv:2509.22611*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale, 2025. *arXiv preprint arXiv:2503.14476*.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. 2025a. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, and 1 others. 2025b. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*.

Sheng Zhang, Qianchu Liu, Guanghui Qin, Tristan Naumann, and Hoifung Poon. 2025a. Medrlvr: Emerging medical reasoning from a 3b base model via reinforcement learning. *arXiv preprint arXiv:2502.19655*.

Yiming Zhang, Yingfan Ma, Yanmei Gu, Zhengkai Yang, Yihong Zhuang, Feng Wang, Zenan Huang, Yuanyuan Wang, Chao Huang, Bowen Song, and 1 others. 2025b. Abench-physics: Benchmarking physical reasoning in llms via high-difficulty and dynamic physics problems. *arXiv preprint arXiv:2507.04766*.

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. 2025. Absolute zero: Reinforced self-play reasoning with zero data. *arXiv preprint arXiv:2505.03335*.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, and 1 others. 2025. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

A Implementation Details

A.1 Token-Level Entropy Calculation

Token-level entropy quantifies the uncertainty of the model's token-generation distribution at a given decoding step. Formally, at decoding step t, the entropy is defined as:

$$\mathcal{H}_t = -\sum_{j=1}^{|V|} p_{t,j} \log p_{t,j}, \tag{6}$$

where $p_t = \pi_\theta(\cdot \mid q, o_{<t}) = \textit{Softmax}(\frac{z_t}{T})$ is the probability distribution over the vocabulary V. Here, π_θ is the model parameterized by θ, q is the

input query, $o_{<t}$ denotes previously generated tokens, z_t represents the pre-softmax logits, and T is the temperature. In this work, the term *token-level entropy* refers specifically to the entropy of the entire vocabulary distribution at a given decoding step, and not to any individual token instance.

Following Cui et al. (2025b), we adopt policy entropy to quantify the predictability or randomness inherent in the actions selected by an agent. Specifically, given policy model π_θ, training dataset $\mathcal{D}$, we measure the average token-level entropy of the policy model on training data, which is defined as follows:

$$\mathcal{H}(\pi_\theta, \mathcal{D}) = -\frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \frac{1}{|o|} \sum_{t=1}^{|o|} \mathbb{E}_{o_t \sim \pi_\theta} \left[\log \pi_\theta(o_t | o_{<t}, q) \right] \tag{7}$$

Such entropy quantifies the uncertainty level of the policy on current prompts and is widely known as sequence-level mean entropy. In practice, we calculate the entropy for each batch of prompts randomly sampled from the training dataset.

A.2 Details of the Entropy Dynamics Similarity Functions

Proposed Similarity Measure via KL Divergence To facilitate the comparison of entropy dynamics with varying lengths, we employ an interpolation-based alignment method. Specifically, given two entropy dynamics τ_i and τ_j (assuming $|\tau_i| < |\tau_j|$), we apply *nearest-neighbor interpolation* to τ_i to align its length with τ_j, then we get the τ_i' with length $|\tau_j|$. Subsequently, to prioritize the fluctuation patterns of token entropy over their absolute magnitudes, we normalize the entropy dynamics using the *Softmax* operation:

$$\hat{\tau}_i = \textit{Softmax}(\tau_i'), \quad \hat{\tau}_j = \textit{Softmax}(\tau_j). \tag{8}$$

Finally, we define the similarity between the two entropy dynamics as:

$$s_{\text{Ours}}(\hat{\tau}_i, \hat{\tau}_j) = -\mathbb{D}_{\text{KL}}(\hat{\tau}_i || \hat{\tau}_j) = -\sum_{t=1}^{|\tau_j|} \hat{\tau}_{i,t} \log \frac{\hat{\tau}_{i,t}}{\hat{\tau}_{j,t}}, \tag{9}$$

where $\mathbb{D}_{\text{KL}}(\cdot, \cdot)$ denotes the KL divergence. Since KL divergence is a non-negative measure of difference where a larger value implies lower similarity, we negate it to ensure that a higher s_{Ours} indicates greater similarity.

In addition to the above implementation adopted in our framework, we explored alternative metrics to verify that the aforementioned similarity measure is relatively optimal. Specifically, we experimented with two other methods for comparison, as described below:

High-entropy Tokens Intersection Similarity The High-entropy Tokens Intersection (HTI) is specialized for entropy dynamics through an analysis of overlap among the key segments that govern exploratory behavior. After aligning two entropy dynamics τ_i and τ_j to a common length N via interpolation, we identify the top 20% of tokens with the highest entropy. Let $\mathcal{I}_i^{top}$ and $\mathcal{I}_j^{top}$ be the sets of indices corresponding to these high-entropy tokens in trajectories. The similarity is computed as:

$$s_{\text{HTI}}(\tau_i, \tau_j) = \sum_{t=1}^{N} \min \left(\tau_{i,t} \cdot \mathbb{I}_{t \in \mathcal{I}_i^{top}}, \tau_{j,t} \cdot \mathbb{I}_{t \in \mathcal{I}_j^{top}} \right), \tag{10}$$

where $\mathbb{I}_{(\cdot)}$ denotes the indicator function. This approach provides an efficient measure of coarse-grained alignment in entropy dynamics, primarily highlighting shared high-uncertainty behaviors. However, it may not capture fine-grained structural details in the lower-entropy segments.

Pearson Linear Similarity The Pearson Linear (PL) similarity characterizes the global linear trend of an entropy dynamic by modeling the relationship between token entropy and its sequential index using linear regression. For each entropy dynamic τ, we extract the slope k_τ of the fitted line and the Pearson correlation coefficient δ_τ. The similarity between two entropy dynamics τ_i and τ_j is defined by integrating their angular proximity and linear consistency:

$$s_{\text{PL}}(\tau_i, \tau_j) = \left| \cos(\arctan k_{\tau_i} - \arctan k_{\tau_j}) \cdot \delta_{\tau_i} \cdot \delta_{\tau_j} \right|, \tag{11}$$

where $\cos(\arctan k_{\tau_i} - \arctan k_{\tau_j})$ quantifies the angular difference between the two fitted lines, while the product $\delta_{\tau_i} \cdot \delta_{\tau_j}$ accounts for the reliability of the linear trends. A significant advantage of the PL similarity is that it does not require length-alignment operations, such as interpolation, maintaining low computational complexity. However, its reliance on linear approximations may limit its ability to capture highly non-linear or local fluctuations within the entropy dynamics.

A.3 Details of Entropy Regularization Baselines

Entropy Loss (He et al., 2025) This method directly incorporates the principles of maximumm

entropy RL by augmenting the GRPO objective with an entropy regularization term. Formally:

$$\mathcal{L}_{\text{AEC}}(\theta) = -\frac{\alpha}{|y_i|\,G}\sum_{i=1}^{G}\sum_{t=1}^{|y_i|}\mathcal{H}_t^i, \qquad (12)$$

where α is a hyperparameter controlling the strength of the entropy regularization, and $\mathcal{H}_t^i$ denotes the token-level entropy at step t for response y_i, as defined in Equation 6. Then the final loss function of Entropy baseline is defined as:

$$\mathcal{L}_{\text{Entropy loss}}(\theta) = \mathcal{L}_{\text{GRPO}}(\theta) + \mathcal{L}_{\text{AEC}}(\theta). \qquad (13)$$

This formulation ensures a lower bound on policy entropy, thereby preserving exploration capability, mitigating entropy collapse, and maintaining learning plasticity during training.

80/20 Strategy (Wang et al., 2025a) Standard GRPO treats all tokens uniformly, often allowing low-entropy tokens to overshadow critical updates. To address this, 80/20 Strategy employs a mask $M_{i,t} = \mathbb{I}(\mathcal{H}_t^i \geq \delta_\gamma)$ to restrict updates to the top γ (e.g., 20%) high-entropy tokens. The objective is formulated as:

$$\mathcal{L}_{80/20}(\theta) = -\frac{1}{N_{HE}}\sum_{i=1}^{G}\sum_{t=1}^{|y_i|} M_{i,t}\min\left(\rho_{i,t}\hat{A}_i, \rho_{i,t}^{\text{clip}}\hat{A}_i\right) + \beta\mathbb{D}_{\text{KL}}(\pi_\theta||\pi_{\text{ref}}), \qquad (14)$$

where $N_{\text{HE}} = \sum_{i,t} M_{i,t}$ denotes the total number of high-entropy tokens in the batch, and $\rho_{i,t}$ is the token-level probability ratio. By prioritizing "forking" tokens that represent pivotal logical transitions, this method prevents the KL regularization from being dominated by trivial linguistic patterns, thereby accelerating convergence and enhancing reasoning robustness.

Clip-Higher (Yu et al., 2025) As a core component of DAPO, it encourages higher entropy by using an asymmetric clipping range in the PPO objective. It addresses the observation that standard symmetric clipping disproportionately restricts probability increases for unlikely tokens. Specifically, the clipping range is decoupled into two parameters: a smaller ϵ_{low} to prevent collapse, and a larger ϵ_{high} to allow greater flexibility for "*exploration tokens*". The clipped ratio in GRPO objective is modified as:

$$\rho_i^{\text{clip-higher}} = \text{clip}(\rho_i, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}). \qquad (15)$$

This strategy mitigates entropy collapse and promotes the generation of more diverse samples by enhancing the policy's exploration capabilities.

Method	Key Hyperparameter	Symbol	Value
Entropy Loss	Entropy coefficient	α	0.001
80/20 Strategy	The fraction of high-entropy tokens	Top-γ	20%
Clip-higher	Upper threshold	$\varepsilon_{\text{high}}$	0.28
	Lower threshold	ε_{low}	0.2
KL-Cov	The fraction of high-covariance	Top-k	0.02%
	KL penalty coefficient	β	1

Table 5: To ensure a solid comparison, we implemented these entropy-based methods using the default hyperparameter configurations specified in their original works, as they are inherently agnostic to dataset scale.

KL-Cov (Cui et al., 2025b) This method counteracts entropy collapse by applying a selective KL-divergence penalty to tokens exhibiting high covariance between their log-probabilities and advantages. Specifically, the strategy selects token indices $\mathcal{I}_{\text{KL}}$ corresponding to the top-k covariance values:

$$\mathcal{I}_{\text{KL}} = \{t \mid \text{Rank}(\text{Cov}(o_t)) \leq k \cdot N_T\}, \qquad (16)$$

where $k \ll 1$ denotes the proportion of tokens targeted for regularization, and N_T is the total number of tokens. For tokens within this set, a KL penalty is imposed to regularize the divergence between the current policy π_θ and the rollout policy $\pi_{\theta_{\text{old}}}$. The resulting policy loss is formulated as:

$$L_{\text{KL-Cov}}(\theta) = \begin{cases} \mathbb{E}_t[\rho_t A_t], & t \notin \mathcal{I}_{\text{KL}} \\ \mathbb{E}_t[\rho_t A_t - \beta\mathbb{D}_{\text{KL}}(\pi_{\theta_{\text{old}}}\|\pi_\theta)], & t \in \mathcal{I}_{\text{KL}} \end{cases} \qquad (17)$$

where the importance sampling ratio ρ_t, A_t is the advantage estimate, and β is a hyperparameter controlling the weight of the KL penalty. By selectively penalizing tokens that exhibit high covariance, KL-Cov constrains the policy update within a stable trust region, thereby preserving exploration diversity and preventing premature convergence.

B Details of Training and Evaluation Datasets

Details of Medicine Datasets MedBullets (Chen et al., 2025a) comprises high-quality multiple-choice questions, providing a rigorous foundation for clinical knowledge. MedXpertQA (Zuo et al., 2025) features expert-level medical queries sourced from professional exams and clinical cases, designed to challenge the model's complex reasoning and domain-specific problem-solving skills. To

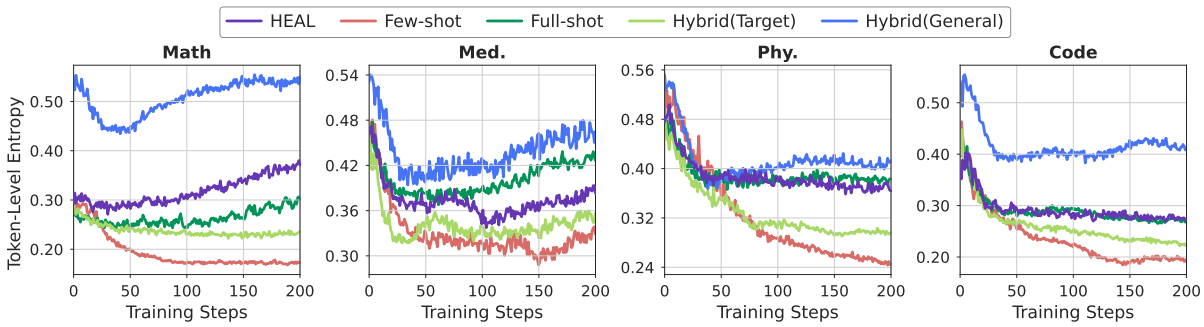

Figure 6: Details of separate average token-level entropy curves on Math, Medicine (Med.), Physics (Phy.), and Code domains. Evaluated on the Qwen3-1.7B-Base model under different datasets settings baselines and our framework HEAL. Hybrid(Target) and Hybrid(General) indicate few-shot RLVR augmented with randomly sampled general-domain data, evaluated on target- and general-domain data.

evaluate expert-level clinical reasoning, we integrate MedBullets and MedXpertQA.

Details of Physics Datasets WebInstruct (Ma et al., 2025) is a large-scale, high-quality instruction-following dataset curated through extensive web mining from Common Crawl. And C-Eval (Huang et al., 2023) is a comprehensive evaluation suite spanning multiple subjects. We specifically employ their physics-related subsets to assess the model's reasoning capabilities across diverse physics problems of varying difficulty levels.

Details of Code Datasets LiveCodeBench (Jain et al., 2024) collects problems from competitive programming platforms with a focus on contamination-free evaluation. We utilize data from versions v1–v4 to construct our training set, while employing LiveCodeBench v5 as an evaluation benchmark. HumanEval Plus (HEval+) (Liu et al., 2023) serves as a rigorous expansion of the original HumanEval dataset (Chen et al., 2021), featuring augmented test cases. We employ them to assess coding capabilities.

Details of Math Datasets Following Wang et al. (2025b), we select 1K examples from DeepScaleR (Luo et al., 2025) as the full-shot dataset for the math domain. For evaluation, we adopt four benchmarks: AMC23 (Art of Problem Solving) comprises 40 challenging problems sourced from prestigious secondary school mathematics competitions. Math500 (Lightman et al., 2023) provides a computationally efficient evaluation through a curated subset of the MATH benchmark's test partition (Verma et al., 2025). The math portion of Olympiad (He et al., 2024) contains international competition problems that demand expert-level logical deduction. Minerva (Lewkowycz et al., 2022) features undergraduate STEM problems from MIT OpenCourseWare designed to evaluate complex multi-step reasoning. We report results using the Avg@4 metric, except for the smaller AMC23 set, which adopts Avg@8.

C Additional Experiment Results

C.1 Entropy Details of Main Results

To provide a more granular perspective on the results presented in Figure 1, we detail the entropy curves across all four target domains, including the Math domain, in Figure 6. Our empirical observations across these diverse domains strongly support the hypothesis that in vanilla low-resource RLVR scenarios, models are highly susceptible to entropy collapse. While standard hybrid training (Hybrid) mitigates this collapse to some extent, the results in Figure 6 demonstrate the superior "*healing*" capability of our HEAL framework. In all four domains, HEAL consistently restores entropy from a collapsed state to levels comparable to the Full-shot baseline. This consistent effectiveness underscores the robust generalization capabilities and stability of the HEAL framework across varying target domains.

C.2 Additional Results on the Math Domain

To further validate the robustness and architectural agnosticism of HEAL within the Math domain, we extended our experiments to encompass various Qwen3 parameter scales and the LLaMA3 architecture. As reported in Table 6, the results are highly consistent with our primary findings: First, HEAL significantly outperforms the Hybrid baseline across all evaluated model scales and architec-

	Data Size		Math				
	$\lvert\mathcal{D}_{tgt}\rvert$	$\lvert\mathcal{D}_{gen}\rvert$	AMC23	Math500	Minerva	Olympiad	Avg.
Qwen3-1.7B-Base			30.94	55.6	13.24	21.48	30.32
↪ Full-shot	1K	0	35.31	**66.2**	19.12	**29.48**	37.53
↪ Few-shot	32	0	34.06	58.4	17.28	23.70	33.36
↪ Only-General	0	10K	29.38	62.8	16.91	26.07	33.79
↪ Hybrid	32	384	32.50	60.2	17.65	24.15	33.63
↪ **HEAL**	32	384	**38.80**	64.4	**21.69**	25.78	**37.67**
Qwen3-4B-Base			42.19	66.2	20.22	34.67	40.82
↪ Full-shot	1K	0	**54.37**	77.4	33.82	38.81	51.10
↪ Few-shot	32	0	50.00	73.4	31.98	37.48	48.22
↪ Only-General	0	10K	47.81	69.2	24.26	34.37	43.91
↪ Hybrid	32	384	51.88	74.4	32.72	38.20	49.30
↪ **HEAL**	32	384	52.81	**77.6**	**35.29**	**39.11**	**51.52**
Qwen3-8B-Base			45.63	63.00	18.01	33.33	39.99
↪ Full-shot	1K	0	**63.75**	78.40	35.66	**42.96**	**55.19**
↪ Few-shot	32	0	58.13	77.20	34.56	41.04	52.73
↪ Only-General	0	10K	53.10	68.60	23.90	37.63	45.81
↪ Hybrid	32	384	54.06	73.80	29.04	39.11	49.00
↪ **HEAL**	32	384	59.69	**78.60**	**38.60**	41.48	54.59
LLaMA3.2-3B-Instruct			25.00	40.8	15.81	13.19	23.70
↪ Full-shot	1K	0	**30.42**	49.4	**22.06**	17.04	**29.73**
↪ Few-shot	32	0	26.88	46.0	19.49	**17.48**	27.46
↪ Only-General	0	10K	25.31	47.6	16.91	17.33	26.79
↪ Hybrid	32	384	26.25	46.8	18.01	16.59	26.91
↪ **HEAL**	32	384	28.44	**49.8**	18.75	17.19	28.55

Table 6: Performance comparison on Math benchmarks. Average scores are computed across the four math datasets. Best results are in bold and the second best are underlined.

	Data Size		Medicine			Physics			Code		
	$\lvert\mathcal{D}_{tgt}\rvert$	$\lvert\mathcal{D}_{gen}\rvert$	MBul.	MedXQA	Avg.	C-Eval	WebIns.	Avg.	HEval+	LCBv5	Avg.
Entropy Loss	32	384	47.70	37.88	42.79	59.92	8.94	34.43	60.98	15.56	38.27
80/20 Strategy	32	384	47.37	37.57	42.47	**62.95**	9.76	36.36	62.19	17.96	40.08
Clip-Higher	32	384	46.38	38.79	42.59	60.49	8.54	34.52	61.59	16.76	39.18
KL-Cov	32	384	48.68	39.31	44.00	61.81	9.35	35.58	60.98	16.16	38.57
HEAL	32	384	**49.34**	**40.00**	**44.67**	62.19	**12.20**	**37.20**	**62.80**	**19.76**	**41.28**

Table 7: More performance comparison between our framework HEAL and other entropy regularization baselines evaluated on the Qwen3-1.7B-Base model under identical training dataset configurations.

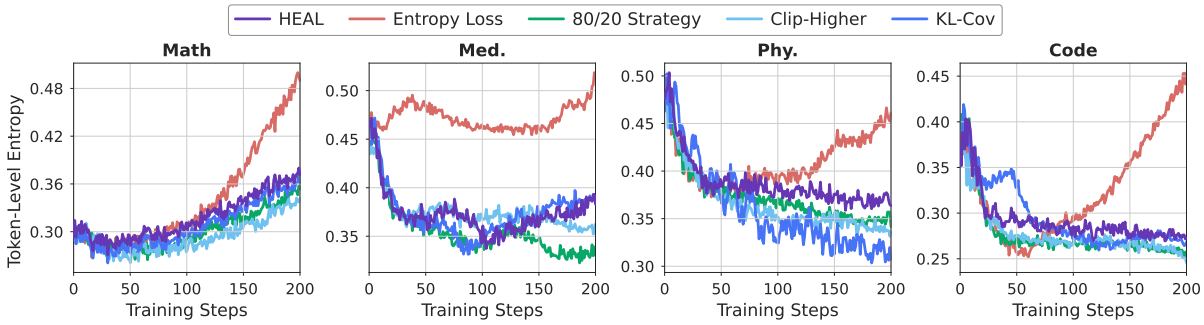

Figure 7: Details of separate average token-level entropy curves on Math, Medicine (Med.), Physics (Phy.), and Code domains. Evaluated on the Qwen3-1.7B-Base model under different entropy regularization methods and our framework HEAL.

tures. Second, HEAL achieves performance that is competitive with, or even superior to, the Full-shot upper bound on several benchmarks. Third, the Only-General baseline often yields results inferior to Few-shot baselines, further highlighting the unique value of HEAL in successfully aligning cross-domain exploration diversity.

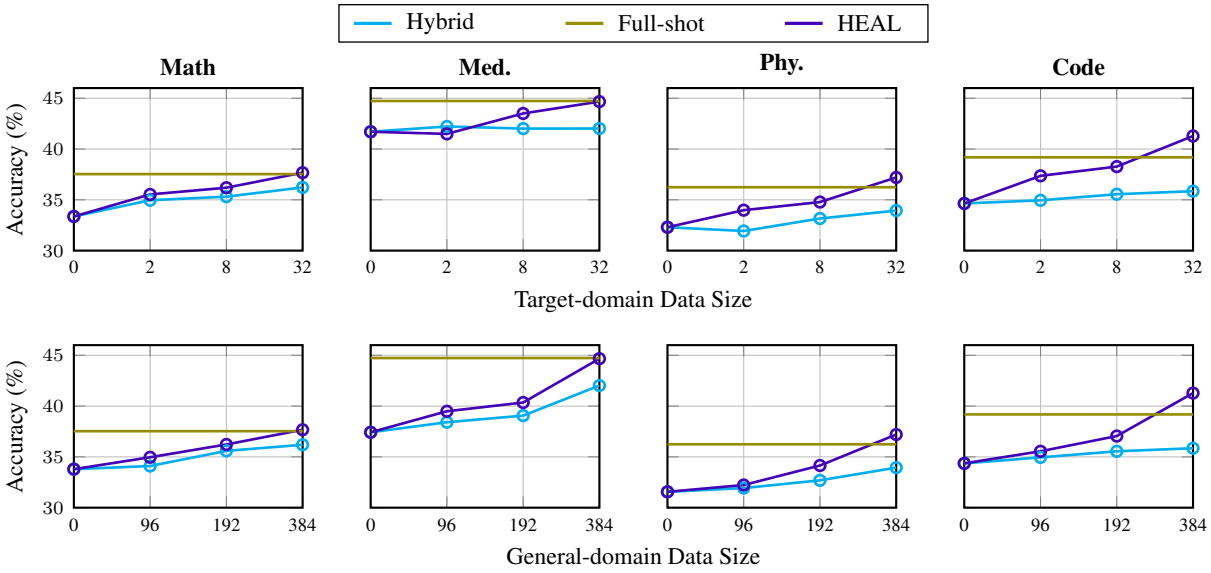

Figure 8: Comprehensive results of the Qwen3-1.7B-Base model across four target domains with respect to different data sizes: (*Top*) Varying target-domain data size while keeping general-domain data fixed; (*Bottom*) Varying general-domain data size while keeping target-domain data fixed.

C.3 More Results of Entropy Regularization Methods

This subsection provides a detailed analysis of the performance on individual benchmarks across all target domains using various methods of entropy regularization. Results in Table 7 show that the standard Entropy Loss baseline leads to poor performance on many tasks. As illustrated in Figure 7, maximizing entropy with a simple scalar-based regularization often causes an entropy explosion, an unintended effect where the model produces disordered and incoherent text due to excessively high entropy, Furthermore, while other competitive methods for entropy regularization keep entropy at moderate levels, their final task performance is still much lower than HEAL. This observation reveals an important insight: high performance in RLVR requires more than just keeping entropy at a certain level; it depends on the dynamics of entropy over time. By aligning the entropy dynamic patterns of the target domain with those of the general-domain, HEAL ensures the model learns when to explore and when to focus on specific logical paths.

C.4 Details about Impact of Data Size

Figure 8 details how performance changes as the amount of data increases across four distinct domains (Math, Code, Medicine, and Physics). Consistent with the analysis in Section 4.3, HEAL exhibits a significantly steeper growth rate than the Hybrid baseline, whether increasing target domain (top row) or general-domain (bottom row) data sizes. Taking the Code and Medicine domains as prime examples, HEAL effectively unlocks the potential of the data, rapidly diverging from the Hybrid and even surpassing the Full-shot upper-bound at specific data scales (e.g., when target domain data size reaches 32).

D Further Analysis

D.1 Exploration Diversity

To investigate the exploration mechanism underlying HEAL, we analyze the model's Entropy Dynamics (EDs) diversity. Mathematically defined in Section 3.2, each ED vector captures the evolution of uncertainty throughout the reasoning process. We hypothesize that the diversity of ED patterns serves as a proxy for the diversity of reasoning paths; specifically, a larger distributional distance between ED sequences indicates that the model is exploring a more heterogeneous set of generation strategies rather than collapsing into a single exploration pattern.

The distance between two EDs is defined as $d = -s_{\text{Ours}}$, based on the similarity metric in Equation 9. Figure 9 visualizes the distances between the EDs of generated samples in a target domain (e.g Math) throughout training. Comparing the baseline vanilla Few-shot RLVR baseline (top) with our HEAL framework (bottom), we observe distinct evolutionary behaviors. Under the

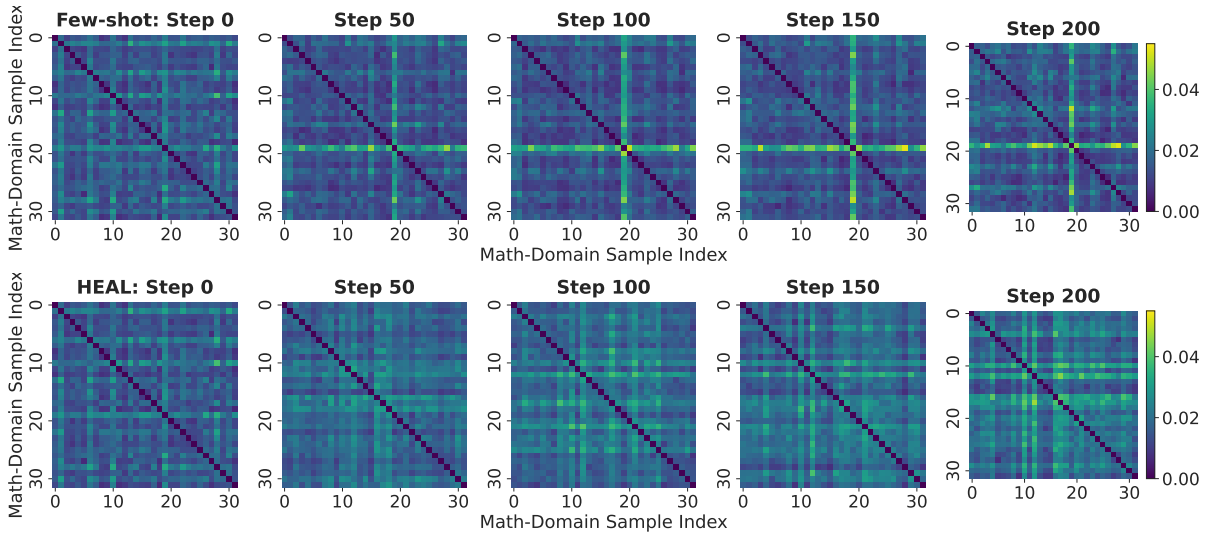

Figure 9: Visualization of Entropy Dynamics (ED) diversity evolution. The heatmaps display the pairwise distances between EDs of generated samples in the Math domain across training steps 0–200. (*Top*): The Few-shot baseline maintains consistently low inter-sample distances (predominantly dark regions), indicating a high degree of homogeneity and limited exploration. (*Bottom*): HEAL exhibits a progressive and widespread increase in distances (shifting to brighter regions), demonstrating that our framework effectively diversifies entropy trajectories.

Few-shot setting, the distances among EDs remain consistently low, indicating a high degree of homogeneity. The dominance of low-distance regions suggests that the model converges to a narrow set of ED patterns, reflecting limited exploration and a tendency toward diversity collapse. In contrast, HEAL exhibits a progressive increase in distances among EDs. This trend indicates that the model under HEAL effectively diversifies its entropy trajectories, thereby mitigating the tendency to converge to few reasoning patterns. This observation validates the effectiveness of our EDA reward. By incorporating soft reward signals derived from general-domain EDs, HEAL encourages the model to maintain higher entropy, effectively unlocking the data's potential by preventing premature convergence to local optima.

D.2 Evaluating Pass@k Metrics on LiveCodeBench

To further evaluate the exploratory potential of HEAL from a discovery perspective, we employ the Pass@k metric on the LiveCodeBench benchmark. As shown in Figure 10, we compared the performance of Qwen3-1.7B-Base and Qwen3-4B-Base models under Few-shot, Full-shot, and HEAL settings, with k varied among 1, 5, 10. A key observation on the Qwen3-4B-Base model is that while the Few-shot baseline performs reasonably at $k = 1$, it lags significantly behind the Full-shot baseline at $k = 10$. This gap suggests that Few-shot models often suffer from *exploration collapse*, where the model over-exploits a few high-probability paths and fails to explore alternative correct solutions. In contrast, HEAL achieves superior performance across all values of k. It maintains high precision at low k while demonstrating an exploratory breadth at high k that rivals or exceeds the Full-shot baseline. This evidence suggests that HEAL successfully demonstrates a superior balance between exploration and exploitation, allowing the model to not only identify the most likely answer but also to maintain the potential to discover deeper reasoning paths in low-resource settings.

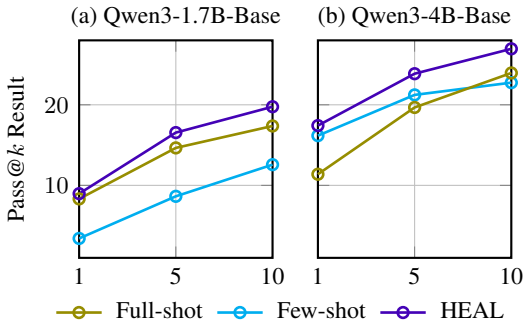

Figure 10: Pass@k results on LiveCodeBench v5 for Qwen3 models. The performance of the 1.7B- and 4B-Base models is compared across Few-shot, Full-shot, and HEAL settings. We investigate the impact of Pass@k by varying k among 1, 5, and 10.