A Survey on *Entropy Mechanism* in Large Reasoning Models

Anonymous ACL submission

Abstract

Entropy mechanism emerges as a key organizing principle for understanding and improving Large Reasoning Models (LRMs). This survey examines how entropy shapes both their training and inference behavior. On the training side, we take a mathematical perspective: casting existing RL-based reasoning algorithms into a unified objective and using this formulation to derive an entropy-centric decomposition of methods, clarifying how different approaches adjust exploration–exploitation "knobs". On the inference side, based on the characteristic of LRM that trading increased inference tokens yields better performance, we summarize methods that leverage entropy to enhance inference performance or reduce uncertainty. Finally, we discuss open challenges and future directions of entropy-driven research for LRMs. Our repository is available on https://anonymous.4open.science/r/Awesome_LRM_with_Entropy-0503.

1 Introduction

Large Reasoning Models (LRMs) (*e.g.*, OpenAI o1 (OpenAI, 2024) and DeepSeek-R1 (Guo et al., 2025)) are designed for slow thinking—deliberate, token-intensive reasoning that markedly improves systematic reasoning and complex problem-solving (Pan et al., 2026; Li et al., 2025d). Their reasoning capabilities are supported by two key components: (i) Reinforcement Learning (RL)–based training, which explicitly encourages and shapes reasoning behavior (Zeng et al., 2025a; Kimi Team, 2025; MiniMax-AI, 2025), and (ii) Test-Time Scaling (TTS) (Muennighoff et al., 2025; Chen et al., 2024b), which allocates substantially more inference-time compute to extend and refine the reasoning trace.

Entropy measures the model's uncertainty over the next-token distribution. A high entropy indicates many possible continuations, whereas a low entropy suggests the model is more certain or conservative. According to existing studies, entropy plays a central role in both the RL training and TTS inference stages of LRMs. During RL, entropy and task reward naturally lie at opposite ends of the exploration–exploitation trade-off (Xue et al., 2025; Zhang et al., 2025c; Audibert et al., 2009; Yue et al., 2025), as illustrated in Fig. 1(left). In settings with sparse rewards, PPO/GRPO-style ratio clipping and batch- or group-mean baselines (Schulman et al., 2017; Shao et al., 2024) induce systematic bias in estimated advantages, driving the policy to concentrate probability mass on a narrow set of high-reward trajectories (Jin et al., 2025; Liu et al., 2025b). This behavior accelerates entropy collapse and premature convergence, sharply reducing exploration and solution-space coverage and ultimately trapping the model in suboptimal policy regions (Cui et al., 2025; Huang et al., 2025b; Hao et al., 2025). During inference, LRMs generate far longer reasoning than conventional Large Language Models (LLMs). The high-entropy segments are typically regarded as an important internal signal and turning points in the reasoning, where it initiates reflection, introduces assumptions, and triggers other key reasoning patterns (Wang et al., 2025f; Shorinwa et al., 2025) as Fig. 1(right). By understanding, controlling, and exploiting the transitions between high- and low-entropy phases

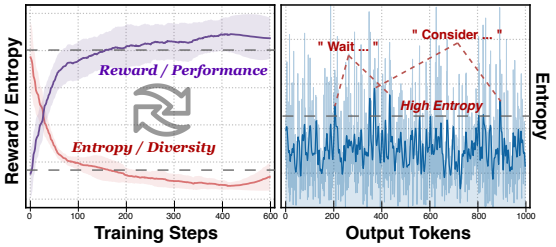

Figure 1: The left panel shows reward and entropy during GRPO training on GSM8K with Qwen3-8B, and the right panel shows entropy at inference after training, where the high entropy parts are shown.

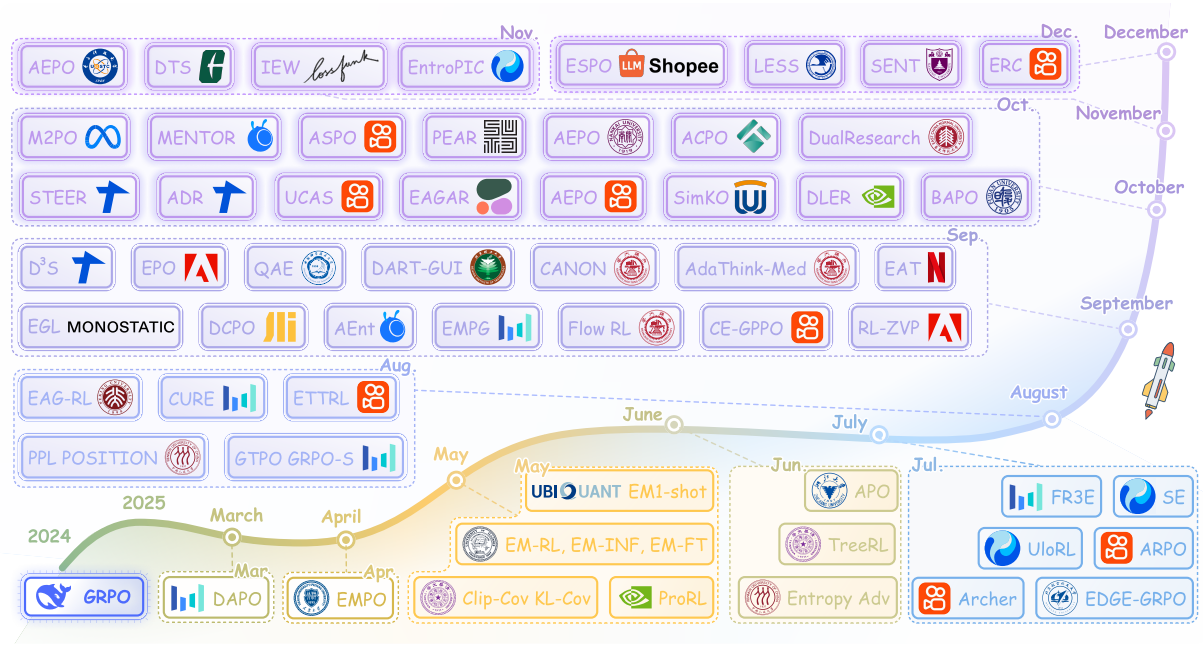

Figure 2: Timeline of papers on entropy mechanisms in LRM within one year after the emergence of GRPO.

during reasoning, researchers can raise the think quality of LRMs (Kayal et al., 2025; Farquhar et al., 2024). Fig. 2 shows the rapid growth of LRM+entropy research, led by top institutes, highlighting entropy's importance in LRMs.

This survey examines entropy mechanisms in LRM training and inference, as shown in Fig. 3, which highlights the relevant work and classifications. On the training side, we introduce a new unifying perspective: we embed existing approaches into a single RL objective and interpret their effects as regularizing specific parameters within that objective. Distinct from prior surveys organized around benchmarks and problem types (Xu et al., 2025a; Liu et al., 2025c), we categorize training methods according to their modified parameters in PPO/GRPO's original optimization objective, explicitly identifying the specific 'knob' that is modified for each. On the inference side, we revisit several recurring criticisms of LRMs (*e.g.*, excessive inference time (Zhang et al., 2025c; Alomrani et al., 2025)) and analyze how entropy-aware TTS techniques leverage intrinsic signals to mitigate these issues in practice. Building on these insights, we highlight key open challenges and outline several promising directions for future research.

2 Entropy Mechanism in LRM Training

LRMs are typically trained with policy-gradient methods, such as PPO and GRPO, to strengthen their inference capabilities (Schulman et al., 2017; Shao et al., 2024; Zhang et al., 2025c). The corresponding training loss is shown in Fig. 4. $\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid q)$ denotes a group of G trajectories generated by the behavior policy $\pi_{\theta_{\text{old}}}$. Each trajectory o_i has a token sequence of length $|o_i|$, and the inner sum runs over token positions $t = 1, \ldots, |o_i|$. $r_{i,t}(\theta)$ is the importance ratio.

Existing training approaches largely operate by adjusting five control knobs: **(i) The sampling term** $q \sim P(Q)$ determines which queries are exposed to the learner and thus which parts of the loss landscape are emphasized during training. **(ii) The advantages of tokens** $\hat{A}_{i,t}$ determine which pattern is encouraged/punished. **(iii) The clip ratio** $\text{clip}(\cdot)$, which controls the magnitude of policy updates by capping the importance ratio $r_{i,t}$. And **(iv) The KL loss** $\beta\, D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})$ regularizes deviation from a reference policy controlled by β. Additionally, some methods directly change **(v) The learning objective itself** $\mathcal{J}_{\text{PPO/GRPO}}(\theta)$, for example by adding or reweighting loss terms. We organize all methods by the parameters they modify to form an entropy-control guideline; methods that affect multiple parameters are presented separately.

2.1 Sampling – $q \sim P(Q)$

Policy learning algorithms learn from self-generated samples, making sampling schemes crucial for training. In GRPO-based methods, within-group sample diversity is essential for effective learning signals. Based on rollout trajectories, sampling approaches can be categorized as: **(1) Single-**

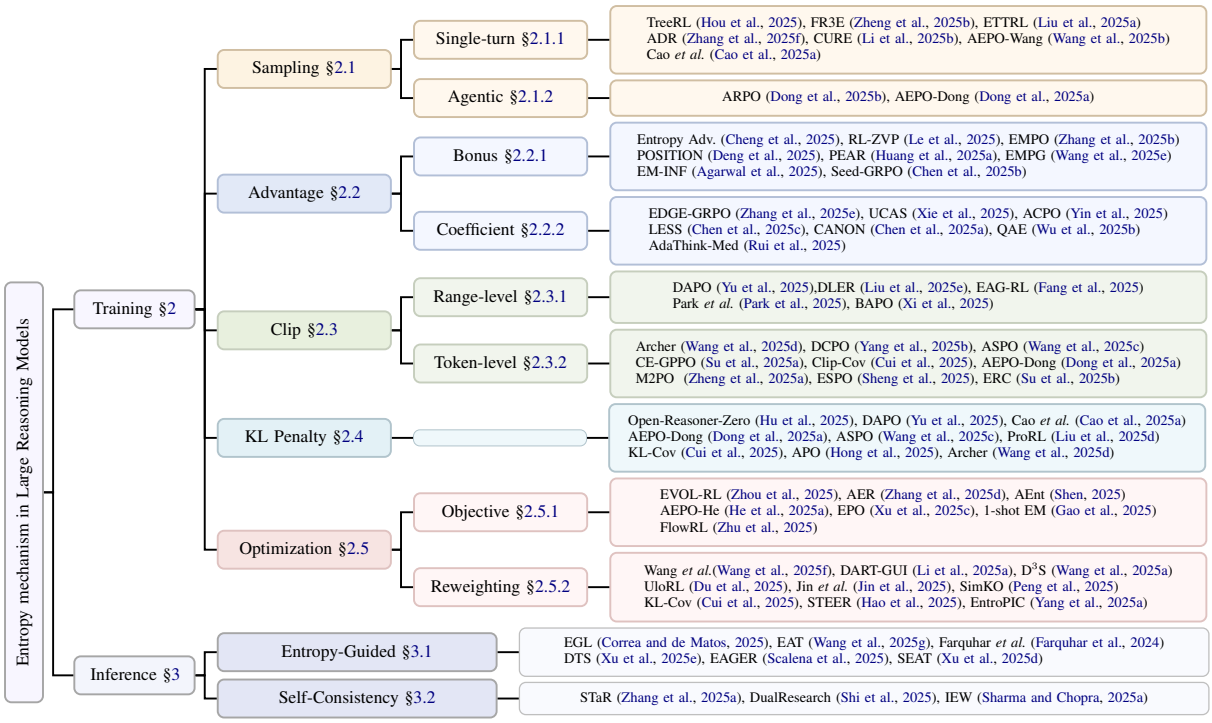

Figure 3: An entropy-centric overview of LRM mechanisms in training and inference.

$$\mathcal{J}_{\text{PPO/GRPO}}(\theta) = \mathbb{E}_{\substack{q\sim P(Q),\\ \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)}} \left[-\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\left(r_{i,t}(\theta)\,\hat{A}_{i,t}, \text{clip}\left(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\right) + \beta\, D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}) \right]$$

Figure 4: The PPO/GRPO optimization objective, where different colors mark different parts.

Turn Sampling, the standard single-round inference commonly used in tasks like math reasoning; **(2) Agentic Sampling**, involving multi-turn rollouts with tool interactions

2.1.1 Single-Turn Sample

AEPO-Wang (Wang et al., 2025b) adopts a dynamic sampling strategy: when the overall entropy is low, it increases the sampling temperature, and when the entropy is high, it decreases the temperature, thereby controlling the diversity of samples. Some methods treat high-entropy tokens in the sampling process as critical turning points and use them as signals to branch the sampling trajectories. TreeRL (Hou et al., 2025) is a representative example: it uses such high-entropy tokens to construct tree-structured sampling results. FR3E (Zheng et al., 2025b), ADR (Zhang et al., 2025f), and ETTRL (Liu et al., 2025a) adopt similar strategies. In contrast, CURE (Li et al., 2025b) emphasizes that the rollout should already be in a high-entropy regime at the very beginning to enhance the diversity of early decisions and concatenates the low-entropy segments of the rollout into the prompt and then re-rolls for new samples. Cao *et al.* (Cao et al., 2025a) directly use entropy as a measure of problem difficulty, constructing a curriculum learning trajectory that progresses from low to high-entropy—that is, from easy to hard.

2.1.2 Agentic Sample

For agentic sampling, the presence of tool-related information that is not directly generated by the model leads to irregular entropy fluctuations. ARPO (Dong et al., 2025b) observes this phenomenon and, analogous to single-turn samples, shares the preceding prefix at high-entropy tool-call turns and then samples new branches from that point. AEPO-Dong (Dong et al., 2025a), as a follow-up, found that multiple consecutive high-entropy rounds would concentrate limited branching resources on a few tracks, causing sampling to "freeze" on a high-entropy chain (high-entropy rollout collapse). Therefore, AEPO-Dong added entropy pre-monitoring and branch penalties to the ARPO framework: on the one hand, adaptively allocating the number of global vs. branch rollouts based on problem entropy/tool entropy; on the other hand, suppressing consecutive high-entropy branches on the same path, distributing exploration

more evenly across different tracks.

2.2 Advantage – $\hat{A}_{i,t}$

In RL, advantage fundamentally influences the direction of optimization, while original GRPO's advantage setting is also considered to be one of the causes of entropy collapse. Some studies directly control entropy changes during the reinforcement process by modifying different advantage calculation methods. They can be categorized from two perspectives: (1) **Entropy as Advantage Bonus**, where entropy or entropy-derived metrics are directly incorporated into the advantage calculation. (2) **Entropy as Coefficient**, where entropy is used to weight the advantage, thereby amplifying or suppressing tokens within specific entropy ranges.

2.2.1 Entropy as Advantage Bonus

A significant body of work incorporates entropy bonus into the advantage function to foster diversity and deeper reasoning. For example, Cheng (Cheng et al., 2025) found that augmenting the advantage function with an entropy term can encourage the LLM to generate longer, deeper reasoning processes, leading to better Pass@K coverage. RL-ZVP (Le et al., 2025) designed a finer-grained reward signal based on policy entropy for zero-variance samples. For fully correct problems, entropy is directly used as the advantage, promoting diverse correct solutions; for fully incorrect problems, the token minus the maximum entropy value is used as the advantage, ensuring that tokens with higher entropy incur smaller penalties, while low-entropy tokens are penalized more, thus encouraging further exploration. Some works (Deng et al., 2025) found that token entropy follows a U-shaped distribution, with higher values at the beginning and end of sequences. In the early stages, high-entropy tokens dominate exploration, while in the terminal stages, high-entropy tokens reflect reasoning uncertainty. Motivated by similar observations, PEAR (Huang et al., 2025a) uses the difference between the entropy of thinking content and answer content as an additional advantage, encouraging more exploration in the early stages and normalization in the later stages.

Unlike the approach of maximizing entropy to encourage exploration in LLMs, EM-INF (Agarwal et al., 2025) demonstrates that minimizing entropy can surprisingly enhance model performance by making the model's outputs more stable. Similarly, EMPG (Wang et al., 2025e) imposes a penalty on future entropy, actively guiding the agent away from high-uncertainty and ambiguous tokens.

Beyond classical entropy, some works note that semantic entropy can also serve as an auxiliary signal in advantage calculations. Seed-GRPO (Chen et al., 2025b) adopts semantic entropy to measure the diversity of generated answers to a prompt, thereby directing the model to focus on samples with higher semantic entropy uncertainty. EMPO (Zhang et al., 2025b) observes that semantic entropy has a strong negative correlation with model accuracy and can be an effective advantage term for minimizing unreliable generations.

2.2.2 Entropy as Advantage Coefficient

Several studies utilize entropy to modulate the importance of tokens or samples during updates. EDGE-GRPO (Zhang et al., 2025e) directly uses $\frac{1}{H(\pi_\theta(\cdot|o))}$ as a coefficient to encourage the update of low-entropy tokens. UCAS (Xie et al., 2025) classified the coefficients for positive and negative sample pairs, encouraging low-entropy for correct samples and high-entropy for incorrect samples. LESS (Chen et al., 2025c) finds that overlap of low-entropy reasoning segments strongly correlates with accuracy, then assigns higher weights to low-entropy segments that recur in correct rollouts while downweighting those that recur in incorrect ones, yielding higher accuracy and more stable reasoning after the RL training. ACPO (Yin et al., 2025) approximated the entropy to calculate the mutual information of each token, i.e., how much additional information a current token brings to the LLM, using this as a coefficient for the advantage to help the model focus on helpful and important steps. Additionally, in algorithms like GRPO, reward normalization within groups is required, and entropy can serve as a classification standard to guide which samples should be calculated.

Beyond direct weighting, entropy also serves as a criterion for advantage grouping. CANON (Chen et al., 2025a) re-groups answers based on entropy to emphasize selecting correct responses with low entropy. QAE (Wu et al., 2025b) proposes a grouping method based on advantage percentage, which prevents both entropy exploration and explosion, providing lower/upper bounds on one-step entropy changes to curb explosion and avoid collapse. Inspired by curriculum learning, AdaThink-Med (Rui et al., 2025) uses entropy as an indicator of problem difficulty, offering different advantage calculations for samples with varying learning efficiency levels.

2.3 Clip Mechanism – $\text{clip}(\cdot)$

The clipping mechanism constrains update magnitude based on $r_{i,t}(\theta)$, creating an approximate trust region that prevents destabilizing large updates on individual tokens. The lower and upper bounds respectively restrict updates that decrease or increase token probabilities, selectively damping gradients and influencing policy entropy. Building on this idea, existing methods can be broadly categorized by control granularity and objectives into two types: **(1) Range-level clip modulation**: directly adjusting the global clipping interval to globally relax or tighten the update range. **(2) Token-level clip modulation**: performing fine-grained control at the token level, assigning different effective clipping behaviors to different types of tokens, or restoring/recalibrating the gradients of tokens.

2.3.1 Range-level Clip Modulation

DAPO (Yu et al., 2025) first discovered that, in practice, the positive samples with low old probabilities often lead to new branches and reflections. Because π_{old} is low, these models are prone to hitting the upper bound and being pruned, potentially limiting the model's ability to learn and explore, causing entropy to collapse rapidly. Therefore, DAPO directly raised the clip upper bound by a constant. DLER's experiments (Liu et al., 2025e) also revealed that the tokens pruned by the upper bound are usually high-entropy tokens worth learning, so it similarly adopted a higher clip upper bound. EAG-RL (Fang et al., 2025) did not use a constant to change the clip range, but adjusted it based on the entropy of each trajectory, giving high-entropy trajectories a higher clip upper bound. Park *et al.* (Park et al., 2025) is even more radical, directly removing the upper bound constraint and raising the lower bound to make error correction less aggressive. While BAPO (Xi et al., 2025) takes into account that off-policy training itself affects clipping: as training progresses, it dynamically raises the upper bound and lowers the lower bound of the clip range, reducing the occurrence of clipping so that the model can keep updating effectively under off-policy conditions.

2.3.2 Token-level Clip Modulation

Since clipping controls probability changes, some works redesign clipping rules to preserve entropy from a probability perspective. Archer (Wang et al., 2025d) distinguishes high-entropy (reasoning) tokens from low-entropy (knowledge) tokens, applying wider clipping bounds to the former and tighter bounds to the latter. DCPO (Yang et al., 2025b) dynamically adjusts clipping based on each token's prior probability: tokens with low old probability receive wider upper bounds, while high-probability tokens receive narrower bounds.

Some methods are designed to modify the core of clip $r_{i,t}(\theta)$. The ratio $r_{i,t}(\theta)$ is flipped (i.e., uses $1/r_{i,t}(\theta)$) by ASPO (Wang et al., 2025c) for tokens with positive advantage, thereby boosting the weight of low-probability tokens while reducing the weight of high-probability tokens, and pairs this with a reversed clipping mechanism to favor tokens that are correct but currently low probability. ESPO (Sheng et al., 2025) also reshapes $r_{i,t}(\theta)$. It recomputes the ratios over different subsequences within the same sample, so that the global importance ratio can be redistributed more precisely to the high-/low-entropy portions of the sequence and also assigns a wider clipping range to the high-entropy segments to encourage exploration. In ERC (Su et al., 2025b), beyond the standard importance ratio $r_{i,t}(\theta)$, the authors introduce an additional entropy ratio that acts as a soft clipping constraint, preventing excessive shifts in policy entropy and thereby stabilizing training. Clip-Cov (Cui et al., 2025) uses covariance to identify situations where high-probability tokens also have high advantages, and then directly clips the gradients of a very small subset of these high-covariance tokens.

Another approach type argues that failing to learn from clipped tokens can allow the model to gradually accumulate biases, thereby accelerating entropy reduction. Both CE-GPPO (Su et al., 2025a) and AEPO-Dong (Dong et al., 2025a) reactivate gradients for clipped tokens. The key difference is that CE-GPPO assigns a reduced learning weight to tokens clipped on either side of the interval, whereas AEPO-Dong only restores tokens whose ratios exceed the upper clipping bound. M2PO (Zheng et al., 2025a) pulls back tokens that would be truncated by clip-high by cutting off only the most extreme high-entropy outliers. It keeps the second moment of importance weights under control while preserving most of the high-entropy tokens even in off-policy settings.

2.4 KL Penalty – $\beta D_{\text{KL}}(\pi_\theta \| \pi_{\textbf{ref}})$

Kullback–Leibler Divergence (KL) is typically formulated as a penalty against a reference policy π_{ref} controlled by β. In short, it's a model exploration limiter based on the reference model, restricting

distribution drift and indirectly regulating/guiding the evolution of policy entropy, preventing excessive entropy collapse. Recent work around this 'knob' largely involves adjusting the KL coefficient $\beta = 0$ through entropy.

Many studies have removed the KL term, i.e., setting β to 0, primarily motivated by the belief that KL regularization itself limits the divergence of exploration, leading to a rapid collapse of entropy, especially when extending the length of the reasoning chain or the number of interaction rounds. There are numerous such studies, so we will only introduce some representative works. From the perspective of traditional RL tasks, early methods proposing the removal of KL include Open-Reasoner-Zero (Hu et al., 2025) and DAPO (Yu et al., 2025); in multi-round agentic tasks, AceReason-Nemetron (Chen et al., 2025d) and MT-GRPO (Zeng et al., 2025b) were among the first to apply this idea. KL-Cov (Cui et al., 2025) set $\beta = 0$ for certain tokens, it identifies high-covariance tokens and applies KL penalties only to these tokens, primarily suppressing those that are already correct and have higher probabilities, preventing them from being further amplified, thereby slowing down the collapse of entropy.

There are also methods that adjust β at the token-level without setting it to zero. Archer (Wang et al., 2025d) and Cao *et al.* (Cao et al., 2025a) reduce the KL regularization strength on high-entropy tokens to promote exploration, while increasing the KL strength on low-entropy tokens to stabilize the model's core capabilities. APO (Hong et al., 2025) retains KL regularization and introduces the factor of sample difficulty in KL control: it weakens the KL constraint for difficult problems to encourage exploration, and strengthens the KL constraint for easier problems to maintain stability and knowledge retention. ProRL (Liu et al., 2025d) periodically replaces the reference policy with the latest snapshot of the online model, effectively recalibrating the KL constraint to maintain exploration and stabilize entropy over the course of long-term RL.

2.5 Optimization Objective – $\mathcal{J}_{\text{PPO/GRPO}}(\theta)$

The optimization objective fundamentally determines the update direction of RL. In LRMs, it governs how the output-token distribution evolves, steering the model toward behaviors aligned with the specified objective. Recent work revisits the PPO/GRPO objective through entropy-aware designs, which can be broadly grouped into two families: **(1) Entropy-Based Optimization Objective**, which allows the model to directly learn an additional entropy-based objective besides the original optimization objective. **(2) Entropy-Guided Loss Reweighting**, which recalculates or reweights the optimization objective to control the update directions applied to different tokens.

2.5.1 Entropy-Based Optimization Objective

A straightforward way to extend the PPO/GRPO objective is to add a global entropy regularizer that increases response diversity. EVOL-RL (Zhou et al., 2025) adopts this idea by adding a negative policy-entropy term, helping maintain mutation rates under label-free, majority-vote reward settings and preventing entropy collapse. Subsequent studies refined this strategy by assigning heterogeneous entropy weights across training stages, samples, or tokens. For instance, AER (Zhang et al., 2025d) introduces an entropy-coefficient loss at the sample level, allocating larger entropy weights to hard samples to encourage exploration while reducing entropy pressure on easy samples that the model already solves reliably.

Instead, AEnt (Shen, 2025) sets entropy coefficients at the token level: it computes a truncated entropy over a high-probability token set to avoid distortions from enforcing uniformity over the long tail. Its adaptive coefficient keeps this truncated entropy within a prescribed range, thus maintaining diversity only within a plausible candidate set. AEPO-He (He et al., 2025a) applies the principle of adaptively controlling policy entropy across reasoning stages to two-stage reasoning in domain-specific settings, enforcing higher entropy during the Reflection phase to mitigate repetitive, mechanical patterns. EPO (Xu et al., 2025c) takes a round-level perspective in the agentic RL, addressing late-stage policy collapse by anchoring each round's entropy to a historical running average, preventing excessive fluctuations across rounds.

Some depart more radically from PPO/GRPO by rewriting the optimization objective directly in terms of entropy. One-shot Entropy Minimization (Gao et al., 2025) discards the PPO/GRPO paradigm entirely and improves reasoning by minimizing the model's conditional entropy on a small set of samples. FlowRL (Zhu et al., 2025) abandons GRPO's reward-maximization objective and devises a GFlowNet-style MSE objective to match the reward-induced distribution, encouraging high reward and high policy entropy at the gradient level.

2.5.2 Entropy-Guided Loss Reweighting

One line of work introduces entropy-based masking to restrict policy-gradient computation to tokens deemed most critical. Wang *et al.* (Wang et al., 2025f) perform a binary split based on local token entropy and retain only high-entropy forking tokens for GRPO updates. DART-GUI (Li et al., 2025a) extends this idea from tokens to multi-round steps by aggregating token entropy into step-level entropy and optimizing decisions only at high-entropy steps. D^{3}S (Wang et al., 2025a) scores each token by the product of its advantage and entropy, selecting the top-K tokens across samples for gradient computation. UloRL (Du et al., 2025) incorporates global entropy trends: in correctly solved samples, it identifies high-confidence tokens and activates a masking mechanism for them when the policy's average entropy falls below a preset threshold.

Other approaches do not discard tokens but instead apply soft reweighting to control the entropy trajectory while preserving most standard learning signals. From an advantage-driven diagnostic viewpoint, Prog-Adv-Reweight (Jin et al., 2025) argues that positive-advantage updates tend to concentrate probabilities and rapidly reduce entropy, whereas negative-advantage updates partially counteract this effect. It therefore temporarily downweights positive-advantage tokens before gradually restoring their influence. SimKO (Peng et al., 2025) refines this idea by first selecting high-entropy tokens and then applying top-K smoothing to distribute positive-advantage signals across a cluster of candidate tokens. For incorrect samples, it further penalizes the most confident wrong token, producing an asymmetric mechanism that both avoids excessive entropy decay and prevents the consolidation of erroneous modes at key decision points. From a coupled entropy–probability statistics viewpoint, Clip-Cov (Cui et al., 2025) measures each token's marginal contribution to entropy change via the covariance between its action probability and the change in logits, and selectively halts gradient backpropagation for a small subset of tokens with high covariance to regulate the rate of entropy decay. STEER (Hao et al., 2025) further explores predicting how much entropy increase or decrease each token update will induce and assigns continuous soft weights accordingly, such that updates expected to cause drastic entropy shifts are strongly downweighted, while milder updates are largely preserved. This fine-grained control smooths the evolution of policy entropy across training. EntroPIC (Yang et al., 2025a) formulates entropy stabilization as a proportional–integral feedback control problem, dynamically adjusting loss coefficients for positive- and negative-advantage terms. By reweighting high-probability tokens, it stabilizes entropy in long-horizon RL training without modifying the PPO/GRPO objective.

3 Entropy Mechanisms in LRM Inference

An increasing number of studies (Nikitin et al., 2024; Agrawal et al., 2024; Jiang et al., 2025) have identified pronounced overthinking phenomena in the generation process of LLMs from different perspectives. Recently, some work introduces entropy into the inference mechanism to both evaluate (Xu and Lu, 2025; Liu et al., 2025f; Nikitin et al., 2024) and regulate (Sharma and Chopra, 2025b; Li et al., 2025c) inference behavior. These explorations collectively reveal a clear trend: **Entropy is evolving from a passive representation of uncertainty into an active control signal for inference**.

Compared with general LLMs, LRMs exhibit stronger dependencies along their reasoning chains and denser branching in path selection, which amplifies performance overhead and error accumulation when inference is not properly controlled. Accordingly, this section focuses on LRM inference scenarios, systematically reviews the key role of entropy during inference, and organizes existing work into two core directions: (i) **Entropy-guided inference** treats entropy as a signal of inference controllability. By monitoring the dynamic evolution of entropy, it adjusts the inference process itself, including determining when to continue, when to pause, and when to trigger error correction, thereby reducing redundant computation and improving inference efficiency. (ii) **Self-consistency enhancement** uses entropy as an important criterion for evaluating inference quality to select higher-quality inference results, thereby improving the reliability and consistency of the final answer.

3.1 Entropy-Guided Inference

Entropy-guided inference establishes an entropy-centric framework that monitors entropy changes in real time to determine how inference should proceed, for example, whether and when to branch during inference is a key factor for inference performance. DTS (Xu et al., 2025e) models the token generation process as a tree and uses the entropy

of the next token to determine when to expand branches. EAGER (Scalena et al., 2025) monitors the top-K entropy of each step to decide whether branching should be triggered, thereby enabling adaptive control over the exploration scope. Meanwhile, researches have begun to explicitly adopt entropy as a criterion for early stopping or regeneration. EAT (Wang et al., 2025g) monitors entropy changes during token-by-token inference to determine when to stop. SEAT (Xu et al., 2025d) defines semantic entropy based on the entropy of semantic clustering distributions across parallel responses, and uses it to decide whether another inference round is necessary. EGL (Correa and de Matos, 2025) computes entropy after the draft pass and generates an uncertainty report to guide adaptive refinement for regeneration when entropy exceeds a threshold, which advances entropy from merely a decision criterion to an inference correction signal.

3.2 Self-Consistency Enhancement

Self-consistency enhancement leverages entropy to assess the credibility of candidate answers or reasoning paths, obtaining more reliable outputs through entropy-aware filtering or weighting. This approach effectively improves LRM output quality, yielding more robust and consistent answers across inference runs. DualResearch (Shi et al., 2025) treats two graphs constructed during inference as posterior answer distributions and uses their Shannon entropy as weighting factors for answer aggregation. Extending self-consistency enhancement to the full inference chain level, IEW (Sharma and Chopra, 2025a) aggregates multiple inference paths based on an inverse-entropy weighting principle, assigning higher weights to more stable, lower-entropy paths. STaR (Zhang et al., 2025a) quantifies uncertainty along the inference chain based on token-level entropy, enabling the selection of the most reliable inference path.

4 Challenge and Future Directions

This survey reviews entropy-related methods for RL training and TTS inference in the LRM era, discussing current limitations and future directions. Notably, all surveyed methods emerged within one year after GRPO, reflecting the rapid advancement of LLMs into the reasoning era. As entropy remains a fundamental challenge, it will continue to drive the research of LRMs. Additionally, Appendix A.1 provides a comprehensive overview of method settings and performance, while Appendix A.2 briefly covers entropy-related issues in earlier pretraining and SFT methods.

Challenge: Entropy to Performance. It is important to emphasize that entropy control is a means, not an end. The goal is to preserve diversity while improving the accuracy of exploration. Although many methods show empirical gains by adjusting entropy, results vary across settings, indicating that the relationship is not a simple trade-off. A central challenge is to deepen our understanding of this link and make entropy-driven improvements more consistently transfer to performance.

Challenge: Intrinsic Relations Between Parameters. Most existing methods focus on adjusting one algorithmic parameter or a small set of parameters, but lack coordination across approaches. There is still a shortage of methods that adopt a global, fine-grained perspective for tuning LRM training and inference. Developing such frameworks would enable a more systematic analysis of how entropy influences learning dynamics and performance.

Future: Entropy in Agentic Models. The deep integration of agentic multi-turn tool calls and reasoning capabilities has greatly enhanced the model's ability to tackle complex problems. This approach is seen as a crucial breakthrough in moving beyond single-turn problem-solving. In the context of agentic tasks, new entropy-related issues arise, particularly in multi-turn tool calls, reasoning, and environmental interaction. Research on dynamically monitoring and controlling policy entropy in such tasks is crucial to avoid instability in training or inference caused by continuous high-entropy decisions or excessive branching. Solving these issues will enable agentic RL to not only "effectively explore" but also "robustly exploit and consolidate".

Future: Entropy in Parallel Thinking. Parallel thinking refers to having the model generate multiple reasoning paths in parallel and then aggregate or synthesize them into a final answer. This offers a simple and efficient way to convert pass@k potential into pass@1 performance. However, it may also amplify current LLM weaknesses in reasoning robustness and stability. By leveraging entropy to monitor the credibility and diversity of these parallel paths, we can filter for high-quality candidates and provide more reliable signals for path fusion. This could help reduce risks such as hallucinations, logical conflicts, and inconsistencies, thereby further improving overall performance.

Limitations

This survey is limited to the LRM era and primarily focuses on entropy-related methods in RL training and inference. Entropy issues in earlier stages of the broader LLM pipeline, such as pretraining and SFT, are only briefly covered in the appendix. Our taxonomy and narrative synthesis necessarily involve subjective choices, which are significant but largely unavoidable in a fast-evolving research landscape. In addition, due to heterogeneous experimental settings across studies (*e.g.*, base models, training data, and hyperparameters), we avoid direct head-to-head comparisons. As a result, some of our observations should be interpreted as descriptive trends rather than definitive conclusions. Finally, given the rapid progress of this field, the coverage and conclusions of this survey may need to be updated as new results emerge, while we try to continue updating new papers on our repository.

References

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

Sudhanshu Agrawal, Wonseok Jeon, and Mingu Lee. 2024. Adaedl: Early draft stopping for speculative decoding of large language models via an entropy-based lower bound on token acceptance probability. In *NeurIPS Efficient Natural Language and Speech Processing Workshop*, pages 355–369. PMLR.

Mohammad Ali Alomrani, Yingxue Zhang, Derek Li, Qianyi Sun, Soumyasundar Pal, Zhanguang Zhang, Yaochen Hu, Rohan Deepak Ajwani, Antonios Valkanas, Raika Karimi, and 1 others. 2025. Reasoning on a budget: A survey of adaptive and controllable test-time compute in llms. *arXiv preprint arXiv:2507.02076*.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. 2009. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902.

Hongye Cao, Zhixin Bai, Ziyue Peng, Boyan Wang, Tianpei Yang, Jing Huo, Yuyao Zhang, and Yang Gao. 2025a. Efficient reinforcement learning with semantic and token entropy for llm reasoning. *arXiv preprint arXiv:2512.04359*.

Yilin Cao, Ruike Zhang, Penghui Wei, Qingchao Kong, and Wenji Mao. 2025b. Perspective-driven preference optimization with entropy maximization for diverse argument generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22479–22496.

Haw-Shiuan Chang, Nanyun Peng, Mohit Bansal, Anil Ramakrishna, and Tagyoung Chung. 2024. Real sampling: Boosting factuality and diversity of open-ended generation via asymptotic entropy. *arXiv preprint arXiv:2406.07735*.

Guanxu Chen, Yafu Li, Yuxian Jiang, Chen Qian, Qihan Ren, Jingyi Yang, Yu Cheng, Dongrui Liu, and Jing Shao. 2025a. Conditional advantage estimation for reinforcement learning in large reasoning models. *arXiv preprint arXiv:2509.23962*.

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*.

Shiqi Chen, Miao Xiong, Junteng Liu, Zhengxuan Wu, Teng Xiao, Siyang Gao, and Junxian He. 2024a. In-context sharpness as alerts: An inner representation perspective for hallucination mitigation. In *International Conference on Machine Learning*, pages 7553–7567. PMLR.

Xinzhu Chen, Xuesheng Li, Zhongxiang Sun, and Weijie Yu. 2025c. Beyond high-entropy exploration: Correctness-aware low-entropy segment-based advantage shaping for reasoning llms. *arXiv preprint arXiv:2512.00908*.

Yang Chen, Zhuolin Yang, Zihan Liu, Chankyu Lee, Peng Xu, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025d. Acereason-nemotron: Advancing math and code reasoning through reinforcement learning. *arXiv preprint arXiv:2505.16400*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

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*.

Yunseon Choi, Sangmin Bae, Seonghyun Ban, Minchan Jeong, Chuheng Zhang, Lei Song, Li Zhao, Jiang Bian, and Kee-Eung Kim. 2024. Hard prompts made interpretable: Sparse entropy regularization for prompt tuning with rl. In *62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, pages 8252–8271. Association for Computational Linguistics (ACL).

Andrew GA Correa and Ana CH de Matos. 2025. Entropy-guided loop: Achieving reasoning through uncertainty-aware generation. *arXiv preprint arXiv:2509.00079*.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, and 1 others. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*.

Jia Deng, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, and Ji-Rong Wen. 2025. Decomposing the entropy-performance exchange: The missing keys to unlocking effective reinforcement learning. *arXiv preprint arXiv:2508.02260*.

Guanting Dong, Licheng Bao, Zhongyuan Wang, Kangzhi Zhao, Xiaoxi Li, Jiajie Jin, Jinghan Yang, Hangyu Mao, Fuzheng Zhang, Kun Gai, and 1 others. 2025a. Agentic entropy-balanced policy optimization. *arXiv preprint arXiv:2510.14545*.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, and 1 others. 2025b. Agentic reinforced policy optimization. *arXiv preprint arXiv:2507.19849*.

Dong Du, Shulin Liu, Tao Yang, Shaohua Chen, and Yang Li. 2025. Ulorl: An ultra-long output reinforcement learning approach for advancing large language models' reasoning abilities. *arXiv preprint arXiv:2507.19766*.

Yue Fang, Yuxin Guo, Jiaran Gao, Hongxin Ding, Xinke Jiang, Weibin Liao, Yongxin Xu, Yinghao Zhu, Zhibang Yang, Liantao Ma, and 1 others. 2025. Toward better ehr reasoning in llms: Reinforcement learning with expert attention guidance. *arXiv preprint arXiv:2508.13579*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Zitian Gao, Lynx Chen, Haoming Luo, Joey Zhou, and Bryan Dai. 2025. One-shot entropy minimization. *arXiv preprint arXiv:2505.20282*.

Andrey Goncharov, Daniil Vyazhev, Petr Sychev, Edvard Khalafyan, and Alexey Zaytsev. 2025. Complexity-aware fine-tuning. *arXiv preprint arXiv:2506.21220*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Zhezheng Hao, Hong Wang, Haoyang Liu, Jian Luo, Jiarui Yu, and 1 others. 2025. Rethinking entropy interventions in rlvr: An entropy change perspective. *arXiv preprint arXiv:2510.10150*.

Chen He, Xun Jiang, Lei Wang, Hao Yang, Chong Peng, Peng Yan, Fumin Shen, and Xing Xu. 2025a. What makes reasoning invalid: Echo reflection mitigation for large language models. *arXiv preprint arXiv:2511.06380*.

Yongquan He, Wenyuan Zhang, Xuancheng Huang, Peng Zhang, Lingxun Meng, Xiang Zhou, Ke Zeng, and Xunliang Cai. 2025b. Don't half-listen: Capturing key-part information in continual instruction tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23649–23668, Vienna, Austria. Association for Computational Linguistics.

Minjie Hong, Zirun Guo, Yan Xia, Zehan Wang, Ziang Zhang, Tao Jin, and Zhou Zhao. 2025. Apo: Enhancing reasoning ability of mllms via asymmetric policy optimization. *arXiv preprint arXiv:2506.21655*.

Zhenyu Hou, Ziniu Hu, Yujiang Li, Rui Lu, Jie Tang, and Yuxiao Dong. 2025. Treerl: Llm reinforcement learning with on-policy tree search. *arXiv preprint arXiv:2506.11902*.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. 2025. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*.

Chen Huang, Wei Lu, and Wenxuan Zhang. 2025a. Pear: Phase entropy aware reward for efficient reasoning. *arXiv preprint arXiv:2510.08026*.

Guanhua Huang, Tingqiang Xu, Mingze Wang, Qi Yi, Xue Gong, and 1 others. 2025b. Low-probability tokens sustain exploration in reinforcement learning with verifiable reward. *arXiv preprint arXiv:2510.03222*.

Jerry Huang, Peng Lu, and Qiuhao Zeng. 2025c. Calibrated language models and how to find them with label smoothing. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 25593–25612. PMLR.

Shuyang Jiang, Yusheng Liao, Ya Zhang, Yanfeng Wang, and Yu Wang. 2025. Overthinking reduction with decoupled rewards and curriculum data scheduling. *arXiv preprint arXiv:2509.25827*.

Renren Jin, Pengzhi Gao, Yuqi Ren, Zhuowen Han, Tongxuan Zhang, and 1 others. 2025. Revisiting entropy in reinforcement learning for large reasoning models. *arXiv preprint arXiv:2511.05993*.

Gautam Siddharth Kashyap, Mark Dras, and Usman Naseem. 2025. Too helpful, too harmless, too honest or just right? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29711–29722.

Aya Kayal, Eduardo Pignatelli, and Laura Toni. 2025. The impact of intrinsic rewards on exploration in reinforcement learning. *Neural Computing and Applications*, pages 1–35.

Jiyeon Kim, Hyunji Lee, Hyowon Cho, Joel Jang, Hyeonbin Hwang, Seungpil Won, Youbin Ahn, Dohaeng Lee, and Minjoon Seo. 2024. Knowledge entropy decay during language model pretraining hinders new knowledge acquisition. *ArXiv*, abs/2410.01380.

Kimi Team. 2025. Kimi k2: Open agentic intelligence. *Preprint*, arXiv:2507.20534.

Thanh-Long V. Le, Myeongho Jeon, Kim Vu, Viet Lai, and Eunho Yang. 2025. No prompt left behind: Exploiting zero-variance prompts in llm reinforcement learning via entropy-guided advantage shaping. *arXiv preprint arXiv:2509.21880*.

Pengxiang Li, Zechen Hu, Zirui Shang, Jingrong Wu, Yang Liu, Hui Liu, Zhi Gao, Chenrui Shi, Bofei Zhang, Zihao Zhang, Xiaochuan Shi, Zedong Yu, Yuwei Wu, Xinxiao Wu, Yunde Jia, Liuyu Xiang, Zhaofeng He, and Qing Li. 2025a. Efficient multi-turn rl for gui agents via decoupled training and adaptive data curation. *arXiv preprint arXiv:2509.23866*.

Qingbin Li, Rongkun Xue, Jie Wang, Ming Zhou, Zhi Li, Xiaofeng Ji, Yongqi Wang, Miao Liu, Zheming Yang, Minghui Qiu, and 1 others. 2025b. Cure: Critical-token-guided re-concatenation for entropy-collapse prevention. *arXiv preprint arXiv:2508.11016*.

Xianzhi Li, Ethan Callanan, Abdellah Ghassel, and Xiaodan Zhu. 2025c. Entropy-gated branching for efficient test-time reasoning. *Preprint*, arXiv:2503.21961.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhiwei Li, Bao-long Bi, Ling-rui Mei, Jun-Feng Fang, Xiao Liang, Zhijiang Guo, and 2 others. 2025d. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.

Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. 2025e. Preserving diversity in supervised fine-tuning of large language models. In *ICLR*.

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

Keliang Liu, Dingkang Yang, Ziyun Qian, Weijie Yin, Yuchi Wang, Hongsheng Li, Jun Liu, Peng Zhai, Yang Liu, and Lihua Zhang. 2025b. Reinforcement learning meets large language models: A survey of advancements and applications across the llm lifecycle. *arXiv preprint arXiv:2509.16679*.

Keliang Liu, Dingkang Yang, Ziyun Qian, Weijie Yin, Yuchi Wang, Hongsheng Li, Jun Liu, Peng Zhai, Yang Liu, and Lihua Zhang. 2025c. Reinforcement learning meets large language models: A survey of advancements and applications across the llm lifecycle. *arXiv preprint arXiv:2509.16679*.

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

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. 2025e. Dler: Doing length penalty right–incentivizing more intelligence per token via reinforcement learning. *arXiv preprint arXiv:2510.15110*.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025f. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6107–6117.

Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Hao Wong, Junbo Niu, Chengyu Shen, Runming He, Yanhao Li, and 1 others. 2025. Learning what reinforcement learning can't: Interleaved online fine-tuning for hardest questions. *arXiv preprint arXiv:2506.07527*.

Yuchun Miao, Sen Zhang, Liang Ding, Yuqi Zhang, Lefei Zhang, and Dacheng Tao. 2025. The energy loss phenomenon in rlhf: A new perspective on mitigating reward hacking. *arXiv preprint arXiv:2501.19358*.

MiniMax-AI. 2025. Minimax-m2. GitHub repository. Accessed: 2025-12-03.

Jisoo Mok, Jaeyoung Do, Sungjin Lee, Tara Taghavi, Seunghak Yu, and Sungroh Yoon. 2023. Large-scale lifelong learning of in-context instructions and how to tackle it. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12573–12589, Toronto, Canada. Association for Computational Linguistics.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. 2025. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20286–20332.

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. 2022. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576.

Dang Nguyen, Ali Payani, and Baharan Mirzasoleiman. 2025. Beyond semantic entropy: Boosting llm uncertainty quantification with pairwise semantic similarity. *arXiv preprint arXiv:2506.00245*.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *Advances in Neural Information Processing Systems*, 37:8901–8929.

OpenAI. 2024. Openai o1 system card. *Preprint*, arXiv:2412.16720.

Qianjun Pan, Wenkai Ji, Yuyang Ding, Junsong Li, Shilian Chen, Junyi Wang, Jie Zhou, Qin Chen, Min Zhang, Yulan Wu, and 1 others. 2026. A survey of slow thinking-based reasoning llms using reinforcement learning and test-time scaling law. *Information Processing & Management*, 63(2):104394.

Jaesung R Park, Junsu Kim, Gyeongman Kim, Jinyoung Jo, Sean Choi, Jaewoong Cho, and Ernest K Ryu. 2025. Clip-low increases entropy and clip-high decreases entropy in reinforcement learning of large language models. *arXiv preprint arXiv:2509.26114*.

Ruotian Peng, Yi Ren, Zhouliang Yu, Weiyang Liu, and Yandong Wen. 2025. Simko: Simple pass@k policy optimization. *arXiv preprint arXiv:2510.14807*.

Yulei Qin, Xiaoyu Tan, Zhengbao He, Gang Li, Haojia Lin, Zongyi Li, Zihan Xu, Yuchen Shi, Siqi Cai, Renting Rui, Shaofei Cai, Yuzheng Cai, Xuan Zhang, Sheng Ye, Ke Li, and Xing Sun. 2025. Learn the ropes, then trust the wins: self-imitation with progressive exploration for agentic reinforcement learning. *arXiv preprint arXiv:2509.22601*.

Shaohao Rui, Kaitao Chen, Weijie Ma, and Xiaosong Wang. 2025. Adathink-med: Medical adaptive thinking with uncertainty-guided length calibration. *arXiv preprint arXiv:2509.24560*.

Daniel Scalena, Leonidas Zotos, Elisabetta Fersini, Malvina Nissim, and Ahmet Üstün. 2025. Eager: Entropy-aware generation for adaptive inference-time scaling. *arXiv preprint arXiv:2510.11170*.

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, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Aman Sharma and Paras Chopra. 2025a. The sequential edge: Inverse-entropy voting beats parallel self-consistency at matched compute. *arXiv preprint arXiv:2511.02309*.

Aman Sharma and Paras Chopra. 2025b. Think just enough: Sequence-level entropy as a confidence signal for llm reasoning. *arXiv preprint arXiv:2510.08146*.

Han Shen. 2025. On entropy control in llm-rl algorithms. *arXiv preprint arXiv:2509.03493*.

Yuepeng Sheng, Yuwei Huang, Shuman Liu, Haibo Zhang, and Anxiang Zeng. 2025. Espo: Entropy importance sampling policy optimization. *arXiv preprint arXiv:2512.00499*.

Jinxin Shi, Zongsheng Cao, Runmin Ma, Yusong Hu, Jie Zhou, Xin Li, Lei Bai, Liang He, and Bo Zhang. 2025. Dualresearch: Entropy-gated dual-graph retrieval for answer reconstruction. *arXiv preprint arXiv:2510.08959*.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*.

Felix Stahlberg and Shankar Kumar. 2025. The role of outgoing connection heterogeneity in feedforward layers of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22487–22495, Suzhou, China. Association for Computational Linguistics.

Alessandro Stolfo, Wes Gurnee, Ben Wu, Yonatan Belinkov, Xingyi Song, Mrinmaya Sachan, and Neel Nanda. 2024. Confidence regulation neurons in language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 125019–125049.

Zhenpeng Su, Leiyu Pan, Minxuan Lv, Yuntao Li, Wenping Hu, Fuzheng Zhang, Kun Gai, and Guorui Zhou. 2025a. Ce-gppo: Controlling entropy via gradient-preserving clipping policy optimization in reinforcement learning.

Zhenpeng Su, Leiyu Pan, Minxuan Lv, Tiehua Mei, Zijia Lin, Yuntao Li, Wenping Hu, Ruiming Tang, Kun Gai, and Guorui Zhou. 2025b. Entropy ratio clipping as a soft global constraint for stable reinforcement learning. *arXiv preprint arXiv:2512.05591*.

Heejae Suh, Yejin Jeon, Deokhyung Kang, Taehee Park, Yejin Min, and Gary Lee. 2025. Enstom: Enhancing dialogue systems with entropy-scaled steering vectors for topic maintenance. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24615–24631.

Benedict Aaron Tjandra, Muhammed Razzak, Jannik Kossen, Kunal Handa, and Yarin Gal. 2024. Fine-tuning large language models to appropriately abstain with semantic entropy. In *Neurips Safe Generative AI Workshop 2024*.

Zhongwei Wan, Hui Shen, Xin Wang, Che Liu, Zheda Mai, and Mi Zhang. 2025a. Meda: Dynamic kv cache allocation for efficient multimodal long-context inference. 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 2485–2497.

Zifu Wan, Ce Zhang, Silong Yong, Martin Q Ma, Simon Stepputtis, Louis-Philippe Morency, Deva Ramanan, Katia Sycara, and Yaqi Xie. 2025b. Only: One-layer intervention sufficiently mitigates hallucinations in large vision-language models. *arXiv preprint arXiv:2507.00898*.

Chao Wang, Tao Yang, Hongtao Tian, Yunsheng Shi, Qiyao Ma, Xiaotao Liu, Ting Yao, and Wenbo Ding. 2025a. Learning more with less: A dynamic dual-level down-sampling framework for efficient policy optimization. *arXiv preprint arXiv:2509.22115*.

Chen Wang, Zhaochun Li, Jionghao Bai, Yuzhi Zhang, Shisheng Cui, Zhou Zhao, and Yue Wang. 2025b. Arbitrary entropy policy optimization: Entropy is controllable in reinforcement finetuning. *arXiv preprint arXiv:2510.08141*.

Jiakang Wang, Runze Liu, Lei Lin, Wenping Hu, Xiu Li, Fuzheng Zhang, Guorui Zhou, and Kun Gai. 2025c. Aspo: Asymmetric importance sampling policy optimization. *arXiv preprint arXiv:2510.06062*.

Jiakang Wang, Runze Liu, Fuzheng Zhang, Xiu Li, and Guorui Zhou. 2025d. Stabilizing knowledge, promoting reasoning: Dual-token constraints for rlvr. *arXiv preprint arXiv:2507.15778*.

Jiawei Wang, Jiacai Liu, Yuqian Fu, Yingru Li, Xintao Wang, Yuan Lin, Yu Yue, Lin Zhang, Yang Wang, and Ke Wang. 2025e. Harnessing uncertainty: Entropy-modulated policy gradients for long-horizon llm agents. *arXiv preprint arXiv:2509.09265*.

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

Xi Wang, James McInerney, Lequn Wang, and Nathan Kallus. 2025g. Entropy after ⟨/Think⟩ for reasoning model early exiting. *Preprint*, arXiv:2509.26522.

Yizhou Wang, Lingzhi Zhang, Yue Bai, Mang Tik Chiu, Zhengmian Hu, Mingyuan Zhang, Qihua Dong, Yu Yin, Sohrab Amirghodsi, and Yun Fu. 2025h. Cautious next token prediction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25685–25697.

Canshi Wei. 2024. Enhancing fine-grained image classifications via cascaded vision language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1857–1871.

Jialiang Wu, Yi Shen, Sijia Liu, Yi Tang, Sen Song, Xiaoyi Wang, and Longjun Cai. 2025a. Improve decoding factuality by token-wise cross layer entropy of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3912–3921.

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

Zhiheng Xi, Xin Guo, Yang Nan, Enyu Zhou, Junrui Shen, Wenxiang Chen, Jiaqi Liu, Jixuan Huang, Zhihao Zhang, Honglin Guo, and 1 others. 2025. Bapo: Stabilizing off-policy reinforcement learning for llms via balanced policy optimization with adaptive clipping. *arXiv preprint arXiv:2510.18927*.

Can Xie, Ruotong Pan, Xiangyu Wu, Yunfei Zhang, Jiayi Fu, Tingting Gao, and Guorui Zhou. 2025. Unlocking exploration in rlvr: Uncertainty-aware advantage shaping for deeper reasoning. *arXiv preprint arXiv:2510.10649*.

Beining Xu and Yongming Lu. 2025. Tecp: Token-entropy conformal prediction for llms. *Mathematics*, 13(20):3351.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, and 1 others. 2025a. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*.

Guowei Xu, Wenxin Xu, Jiawang Zhao, and Kaisheng Ma. 2025b. Weft: Weighted entropy-driven finetuning for dllms. *arXiv preprint arXiv:2509.20863*.

Wujiang Xu, Wentian Zhao, Zhenting Wang, Yu-Jhe Li, Can Jin, Mingyu Jin, Kai Mei, Kun Wan, and Dimitris N. Metaxas. 2025c. Epo: Entropy-regularized policy optimization for llm agents reinforcement learning. *arXiv preprint arXiv:2509.22576*.

Zenan Xu, Zexuan Qiu, Guanhua Huang, Kun Li, Siheng Li, Chenchen Zhang, Kejiao Li, Qi Yi, Yuhao Jiang, Bo Zhou, and 1 others. 2025d. Adaptive termination for multi-round parallel reasoning: An universal semantic entropy-guided framework. *arXiv preprint arXiv:2507.06829*.

Zicheng Xu, Guanchu Wang, Yu-Neng Chuang, Guangyao Zheng, Alexander S Szalay, Zirui Liu, and Vladimir Braverman. 2025e. Dts: Enhancing large reasoning models via decoding tree sketching. *arXiv preprint arXiv:2511.00640*.

Boyang Xue, Fei Mi, Qi Zhu, Hongru Wang, Rui Wang, Sheng Wang, Erxin Yu, Xuming Hu, and Kam-Fai Wong. 2025. Ualign: Leveraging uncertainty estimations for factuality alignment on large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6002–6024.

Kai Yang, Xin Xu, Yangkun Chen, Weijie Liu, Jiafei Lyu, Zichuan Lin, Deheng Ye, and Saiyong Yang. 2025a. Entropic: Towards stable long-term training of llms via entropy stabilization with proportional-integral control. *arXiv preprint arXiv:2511.15248*.

Shihui Yang, Chengfeng Dou, Peidong Guo, Kai Lu, Qiang Ju, Fei Deng, and Rihui Xin. 2025b. Dcpo: Dynamic clipping policy optimization. *arXiv preprint arXiv:2509.02333*.

Yao Yao, Zuchao Li, and Hai Zhao. 2024. Sirllm: Streaming infinite retentive llm. In *Proceedings of the 62nd Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 2611–2624.

Junxi Yin, Haisen Luo, Zhenyu Li, Yihua Liu, Dan Liu, Zequn Li, and Xiaohang Xu. 2025. Pinpointing crucial steps: Attribution-based credit assignment for verifiable reinforcement learning. *arXiv preprint arXiv:2510.08899*.

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

Xiaojian Yuan, Tianyu Pang, Chao Du, Kejiang Chen, Weiming Zhang, and Min Lin. 2024a. A closer look at machine unlearning for large language models. *arXiv preprint arXiv:2410.08109*.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Shengping Liu, Jun Zhao, and Kang Liu. 2024b. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3903–3922.

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

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, and 1 others. 2025a. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.

Siliang Zeng, Quan Wei, William Brown, Oana Frunza, Yuriy Nevmyvaka, Yang Katie Zhao, and Mingyi Hong. 2025b. Reinforcing multi-turn reasoning in llm agents via turn-level credit assignment. In *ICML 2025 Workshop on Computer Use Agents*.

Huajian Zhang, Mingyue Cheng, Yucong Luo, and Xiaoyu Tao. 2025a. Star: Towards cognitive table reasoning via slow-thinking large language models. *arXiv preprint arXiv:2511.11233*.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. 2025b. Right question is already half the answer: Fully unsupervised llm reasoning incentivization. *arXiv preprint arXiv:2504.05812*.

Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Wenyue Hua, Haolun Wu, Zhihan Guo, Yufei Wang, Niklas Muennighoff, and 1 others. 2025c. A survey on test-time scaling in large language models: What, how, where, and how well? *arXiv preprint arXiv:2503.24235*.

Xiaoyun Zhang, Xiaojian Yuan, Di Huang, You Wang, Hu Chen, Jingqing Ruan, Kejiang Chen, and Xing Hu. 2025d. Rediscovering entropy regularization: Adaptive coefficient unlocks its potential for llm reinforcement learning. *arXiv preprint arXiv:2510.10959*.

Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. 2025e. Edge-grpo: Entropy-driven grpo with guided error correction for advantage diversity. *arXiv preprint arXiv:2507.21848*.

Yujian Zhang, Keyu Chen, Zhifeng Shen, Ruizhi Qiao, and Xing Sun. 2025f. Adaptive dual reasoner: Large reasoning models can think efficiently by hybrid reasoning. *arXiv preprint arXiv:2510.10207*.

Zhaohan Zhang, Ziquan Liu, and Ioannis Patras. 2025g. Get confused cautiously: Textual sequence memorization erasure with selective entropy maximization. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10924–10939.

Haizhong Zheng, Jiawei Zhao, and Beidi Chen. 2025a. Prosperity before collapse: How far can off-policy rl reach with stale data on llms? *arXiv preprint arXiv:2510.01161*.

Tianyu Zheng, Tianshun Xing, Qingshui Gu, Taoran Liang, Xingwei Qu, Xin Zhou, Yizhi Li, Zhoufutu Wen, Chenghua Lin, Wenhao Huang, and 1 others. 2025b. First return, entropy-eliciting explore. *arXiv preprint arXiv:2507.07017*.

Yujun Zhou, Zhenwen Liang, Haolin Liu, Wenhao Yu, Kishan Panaganti, Linfeng Song, Dian Yu, Xiangliang Zhang, Haitao Mi, and Dong Yu. 2025. Evolving language models without labels: majority drives selection, novelty promotes variation. *arXiv preprint arXiv:2509.15194*.

Wenhong Zhu, Hongkun Hao, Zhiwei He, Yiming Ai, and Rui Wang. 2024. Improving open-ended text generation via adaptive decoding. In *Proceedings of the 41st International Conference on Machine Learning*, pages 62386–62404.

Xuekai Zhu, Daixuan Cheng, Dinghuai Zhang, Hengli Li, Kaiyan Zhang, Che Jiang, Youbang Sun, Ermo Hua, Yuxin Zuo, Xingtai Lv, Qizheng Zhang, Lin Chen, Fanghao Shao, Bo Xue, Yunchong Song, Zhenjie Yang, Ganqu Cui, Ning Ding, Jianfeng Gao, and 4 others. 2025. Flowrl: Matching reward distributions for llm reasoning. *arXiv preprint arXiv:2509.15207*.

A Appendix

A.1 Settings & Results Of Entropy-related Methods

We observe that existing methods are rarely compared head-to-head in a systematic manner, and the community's widely accepted and commonly used baselines are typically GRPO or DAPO. Meanwhile, differences in base models, training data, and hyperparameter choices can substantially affect RL training outcomes. With many contemporaneous studies emerging in parallel, these factors further complicate fair evaluations of the effectiveness of different approaches. To mitigate this confounding issue, this survey organizes methods by base model and annotates the detailed training setups reported for most works in Tab. 1, Tab. 2. Our goal is to provide researchers with a comprehensive visual overview rather than an authoritative comparison or ranking of methods. All configuration details are sourced from the original papers and linked repositories. Meanwhile, this paper compiles and summarizes the acquisition addresses of mainstream base models employed in the experiments, covering models such as DeepSeek-R1-Distill-Qwen https://huggingface.co/deepseek-ai, LLaMA https://huggingface.co/meta-LLaMA and Qwen https://huggingface.co/Qwen.

To facilitate a clear evaluation of the improvements in model reasoning capabilities achieved by different methods, the following clarifications are provided. Methods that were not evaluated on mathematical benchmarks are excluded from our analysis, such as EMPG (Wang et al., 2025e), EPO (Xu et al., 2025c), and EPPO (Miao et al., 2025). Furthermore, given the discrepancies in the benchmarks adopted by different methods, we only select the metrics from mathematical benchmarks, namely AIME24, AIME25, MATH500, Olympiad, Minerva, MATH, and AMC23, for statistical analysis, in accordance with the frequency of benchmark utilization. Benchmarks with low evaluation frequency (e.g., AMC, GSM8K, AMC24) as well as non-mathematical benchmarks (e.g., LiveCodeBench, CoQA, WebWalker) are eliminated from the statistics. Correspondingly, to ensure that all methods are included in the statistical scope, a small number of base models (e.g., BP-Math-32B, GLM-9B) are also excluded. Regarding the challenge of standardizing evaluation metrics across models, two types of tables are constructed in this study: Tab. 1 only includes methods evaluated by the pass@1 and accuracy (acc) metrics; Tab. 2 covers all types of evaluation metrics, with specific metric names labeled prior to the corresponding values. For the statistics of experimental parameters, this study focuses on the parameter configurations in the training phase. For methods involving multi-stage training, parameter variations across stages are denoted in the format of "8k/16k". The training datasets are restricted to mathematical domains. Additionally, "MATH (3–5)" denotes the subset of the MATH dataset with difficulty levels ranging from 3 to 5.

Specifically, we systematically extract and document the following key training hyperparameters from each paper. We record the "Global Size" (total batch size across all devices for one parameter update) and "Mini Batch" (per-device batch size), which determine the gradient computation granularity and memory requirements during training. The "Learn. Rate" (learning rate) controls the step size of parameter updates and is critical for training stability and convergence speed. "Data Col." (data columns) refers to the number of data fields used in training. "Train. Steps" indicates the total number of parameter update iterations, representing the complete training trajectory from initialization to convergence. "Rollout Num." specifies the number of candidate response sequences generated per prompt during policy optimization, which directly impacts sample efficiency and exploration diversity. "Samp. Temp." (sampling temperature) modulates the randomness of generated outputs, with higher values increasing diversity. "Max Len." (maximum response length) defines the upper bound of generated sequence lengths.

A.2 Entropy in the Broader LLM Training and Inference Pipeline

Although our survey focuses on entropy mechanisms specifically for LRMs, entropy-driven techniques have also been explored throughout the broader LLM pipeline, including pretraining, supervised finetuning, and decoding. To maintain conceptual completeness while preserving narrative focus in the main text, we provide in this appendix a concise overview of such methods. These works are not explicitly designed for LRMs, yet they are conceptually aligned with the entropy-centric framework developed in this survey. As shown in Fig. 5, we categorize the related methods into three groups along the LLM pipeline: pretraining, supervised fine-tuning (SFT), and decoding,

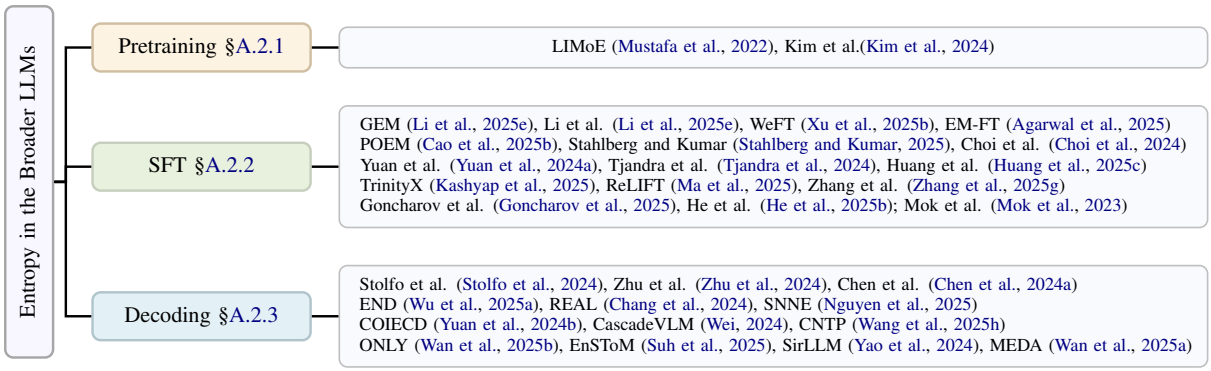

Figure 5: An entropy-centric overview of LMMs.

and we list representative works for each stage.

A.2.1 Entropy in Pretraining

During the pretraining stage, entropy mechanisms are leveraged to regulate knowledge organization and expert utilization within large models. LIMoE (Mustafa et al., 2022) imposes dual entropy constraints on the routing distribution: the minimum-entropy loss encourages each token to confidently select a single expert, improving determinism and computational efficiency, while the maximum-entropy loss promotes broader expert activation at the sample level, preserving representation diversity. Researchers also introduce the concept of knowledge entropy (Kim et al., 2024), demonstrating that knowledge entropy gradually decreases as pretraining progresses, making it increasingly difficult for the model to acquire new knowledge and exacerbating catastrophic forgetting.

A.2.2 Entropy in Supervised Fine-Tuning

During the SFT stage, entropy mechanisms play an important role in the training process in multiple ways. Entropy is widely used to improve reasoning ability after SFT. GEM (Li et al., 2025e) introduces maximum-entropy regularization during fine-tuning, encouraging the model to maintain higher entropy and avoiding concentrating high probability on a single token, thereby improving reasoning performance. WeFT (Xu et al., 2025b) increases the training weight of high-entropy regions in diffusion models, so that more model capacity is allocated to reasoning-critical steps. EM-FT (Agarwal et al., 2025) minimizes token-level distribution entropy during fine-tuning, making the model more inclined to select more reliable steps when generating chains of reasoning. POEM (Cao et al., 2025b) encourages viewpoint diversity in generated semantic representations via entropy maximization. Another work proposes a new fine-tuning loss (Stahlberg and Kumar, 2025) that improves fine-tuning by reducing the outgoing connection entropy of FFN layers. In addition, sparse Tsallis entropy regularization (Choi et al., 2024) filters low-probability tokens, enabling fine-tuning to learn more interpretable hard prompts.

Entropy plays an important role in reducing hallucinations after SFT. The study (Yuan et al., 2024a) incorporates a maximum-entropy mechanism into SFT, increasing predictive entropy on the forget set and thereby reducing hallucinations. Another study (Tjandra et al., 2024) uses semantic entropy as an unlabeled, data-driven signal to fine-tune large language models, so that they proactively refuse when uncertain and reduce hallucination risk. Label smoothing (Huang et al., 2025c) is also introduced into LLM fine-tuning and is explained by its equivalent KL regularization toward the uniform distribution, thereby reducing hallucinations.

Entropy also stabilizes the SFT training process and prevents collapse. TrinityX (Kashyap et al., 2025) adds an entropy regularizer in router training, encouraging more diverse expert selection while avoiding excessive collapse to a small number of experts. ReLIFT (Ma et al., 2025) incorporates an entropy regularization term into the online SFT objective to stabilize training and mitigate distribution collapse. Another work uses entropy as an objective by maximizing next-token distribution entropy on the forgetting set, mitigating privacy risks without significantly harming general capabilities (Zhang et al., 2025g).

Entropy is also used for SFT data processing. A study uses the model's single-token answer entropy to measure question difficulty, partitions data into different difficulty levels, and applies different training strategies for different levels (Goncharov

et al., 2025). Other works use entropy for sample filtering and data replay during fine-tuning to improve training efficiency or preserve key capabilities (He et al., 2025b; Mok et al., 2023).

Overall, entropy is not only a metric for characterizing LLMs' uncertainty, but also an efficient signal that directly guides SFT optimization, affecting key components such as training objective design, weight allocation, and sample selection; therefore, it plays an important role in SFT.

A.2.3 Entropy in Decoding

In the decoding of LLMs, entropy serves as a key control signal for decoding, effectively improving the stability and consistency of generation. The study (Stolfo et al., 2024) finds that neurons inside the model influence generative behavior by regulating their distribution entropy. Therefore, the study (Zhu et al., 2024) continuously monitors changes in entropy before and after token generation to promptly correct the generation direction when unstable fluctuations occur. Meanwhile, entropy is also injected as a penalty term into probability adjustment (Chen et al., 2024a), making the model more inclined to select low-entropy, high-confidence tokens. Furthermore, END (Wu et al., 2025a) incorporates cross-layer entropy to enable more consistent risk estimation across different feature levels. REAL (Chang et al., 2024) computes residual entropy that helps the model adaptively shrink or expand the sampling space during decoding. In long-text scenarios, SNNE (Nguyen et al., 2025) introduces continuous semantic similarity to unify and extend the semantic entropy framework, improving hallucination detection.

Entropy is further employed as a mechanism for scheduling and regulating the inference process. COIECD (Yuan et al., 2024b) determines whether a knowledge conflict has occurred by assessing the information-entropy constraint of the token, triggering a more cautious sampling strategy. CascadeVLM (Wei, 2024) uses entropy to measure the uncertainty of the sample and routes different samples accordingly. CNTP (Wang et al., 2025h) relies on the predictive entropy of the current token to determine whether multiple sampling should be performed, improving inference quality and reliability. At the attention level, entropy can also serve as a trigger signal. ONLY (Wan et al., 2025b) calculates the Text-to-Visual Entropy Ratio for each layer to identify target layers for attention intervention. EnSToM (Suh et al., 2025) utilizes the internal entropy of the LLM as an uncertainty signal to dynamically control activation steering, thereby improving the service capabilities of the dialogue system.

Furthermore, in long dialogues or multimodal tasks, entropy begins to guide the dynamic management of context and resources. SirLLM (Yao et al., 2024) leverages token entropy and memory decay mechanisms to dynamically filter and retain the most important KV cache, enhancing the model's ability to handle long dialogues. MEDA (Wan et al., 2025a) uses cross-modal attention entropy to determine the size of the KV cache for each layer, thereby achieving efficient multimodal inference.

Table 1: Performance of Different Entropy-Based Methods on Mathematical Benchmarks Evaluated by Pass@1 and Accuracy

Base Model	Method	AIME24	AIME25	MATH500	Olympiad	Minerva	MATH	AMC23	open-source	Train Data	Global Size	Mini Batch	Learn. Rate	Data Col.	Train. Steps	Rollout Num.	Samp. Temp.	Max Len.
DeepSeek-R1-Distill-Qwen Models																		
DeepSeek-R1-Distill-Qwen-1.5B	PEAR (Huang et al., 2025a)	23.33	–	77.20	–	–	–	70.00	✓	GSM8K	128	-	1E-6	7473	-	8	0.6	16K
	DA-DLER-R1-1.5B (Liu et al., 2025e)	34.37	–	–	48.70	44.89	86.70	–	✓	DeepScaleR-Preview	512	64	1E-6	40K	600	16	1.0	4000
	DLER-R1-1.5B (Liu et al., 2025e)	34.38	–	–	48.31	43.59	86.95	–	✓	DeepScaleR-Preview	512	64	1E-6	40K	450	16	1.0	4000
	ProRL (Liu et al., 2025d)	48.13	33.33	–	60.22	47.98	91.89	–	✓	DeepScaleR	256	64	2E-6	40k	-	16	1.2	8k
	ADR (Zhang et al., 2025f)	36.5	23.3	81.0	–	–	–	–	✗	OpenMathReasoning, DeepScaleR-Preview	-	-	-	-	-	-	-	8k/16k
DeepSeek-R1-Distill-Qwen-7B	TreeRL (Hou et al., 2025)	60.8	–	94.4	57.1	–	–	–	✓	MATH-train, NuminaMath	480	-	1.5E-6	-	-	30	1.2	8K
	DA-DLER-R1-7B (Liu et al., 2025e)	53.9	–	–	61.16	53.60	94.17	–	✓	DeepScaleR-Preview	512	64	1.0E-6	40K	600	16	1.0	4000
	DA-DLER-R1-7B (Liu et al., 2025e)	55.62	–	–	60.48	53.88	94.21	–	✓	DeepScaleR-Preview	512	64	1.0E-6	40K	450	16	1.0	4000
LLaMA Models																		
LLaMA-3.1-8B	TTRL (Liu et al., 2025a)	10	-	63.7	-	-	-	-	✗	-	-	-	-	-	-	-	0.6	3K
	ETMR (Liu et al., 2025a)	16.9	-	59.5	-	-	-	-	✗	AIME24, AMC, MATH-500	-	-	-	-	-	32	0.6	3K
LLaMA-3.1-8B-Instruct	AEPO-Dong (Dong et al., 2025a)	26.7	16.7	65.8	-	-	80.6	-	✓	-	128	16	-	-	-	16	0.6	20K
	ARPO (Dong et al., 2025b)	23.3	16.7	64.6	-	-	80.2	-	✓	Tool-Star, STILL	128	16	-	-	-	-	-	4K
	EM_INF (Agarwal et al., 2025)	3.3	-	43.0	16.4	22.8	-	-	✓	-	-	-	-	-	-	-	0.1	-
	GRPO + D^{3}S (Wang et al., 2025a)	5.3	0.1	-	3.3	22.5	35.9	20.3	✗	DeepScaleR	-	-	5E-7	-	-	32	0.6	-
LLaMA-3.2-3B-Instruct	SimKO (Peng et al., 2025)	13.8	1.0	54.6	21.0	18.5	-	35.2	✓	GSM8K, MATH	1K	256	1E-6	-	-	8	1.0	-
Qwen-2.5 Models																		
Qwen-2.5-3B	TTRL (Liu et al., 2025a)	7.9	-	72.2	-	-	-	40.7	✗	AIME24, AMC, MATH-500	-	-	-	-	-	-	0.6	3K
	ETMR (Liu et al., 2025a)	9.2	-	71.7	-	-	-	-	✗	AIME24, AMC, MATH-500	-	-	-	-	-	-	-	-
Qwen-2.5-7B	SimKO (Peng et al., 2025)	16.3	9.4	76.7	38.7	35.2	-	-	✓	MATH	1K	256	1E-6	-	-	8	1.0	-
	FR3E (Zheng et al., 2025b)	25.2	-	79.0	42.1	39.0	-	-	✓	DeepScaler, SimpleRL (3-5)	512	128	1E-6	-	-	16	-	16k
Qwen-2.5-7B-Instruct	AEPO-Dong (Dong et al., 2025a)	33.3	33.0	80.4	-	-	90.0	-	✓	-	128	16	-	1K	-	16	0.6	20K
	ARPO (Dong et al., 2025b)	30	30.0	78.8	-	-	88.8	-	✓	Tool-Star, STILL	128	-	-	-	-	-	-	4K
	EM_INF (Agarwal et al., 2025)	11.1	-	73.8	38.2	41.2	-	45.8	✓	-	-	-	-	-	-	-	0.1	-
Qwen-2.5-14B	TreeRL (Hou et al., 2025)	20.8	-	81.7	44.6	-	-	55.9	✓	MATH-train, NuminaMath	480	-	1.5E-6	-	-	30	1.2	8K
	DCPO (Yang et al., 2025b)	20	85.0	84.6	-	-	-	-	✓	MATH-DAPO-17k, MATH (3-5)	512	32	-	25k	400	-	1.0	3K
Qwen-2.5-32B	FR3E (Zheng et al., 2025b)	40.2	-	87.4	51.7	45.6	-	-	✓	DeepScaler, SimpleRL (3-5)	512	128	1E-6	-	-	16	-	16k
Qwen-2.5-Math Models																		
Qwen-2.5-Math-1.5B	GRPO + D^{3}S (Wang et al., 2025a)	11.2	6.9	-	10.6	23.5	53.7	48.6	✗	DeepScaleR	-	-	5E-7	-	-	32	1.0	-
	GSPO + D^{3}S (Wang et al., 2025a)	11.4	8.2	-	10.5	22.9	54.0	48.4	✗	DeepScaleR	-	-	5E-7	-	-	32	1.0	-
	TTRL (Liu et al., 2025a)	15.8	-	73.0	-	–	–	–	✗	AIME24, AMC, MATH-500	-	-	-	-	-	-	0.6	3K
	EDGE-GRPO (Zhang et al., 2025e)	10	-	73.20	37.33	29.04	–	–	✓	DeepScaleR-Random-1K, DeepScaleR-Hard-1K	-	-	1E-6	2k	1000	-	-	1K
	EMPO (Zhang et al., 2025b)	13.3	-	-	36.6	32.4	73.0	55.0	✓	-	–	–	–	–	–	–	–	–
	UCAS (Xie et al., 2025)	23.3	-	80.6	42.1	31.6	–	–	✓	MATH (3-5)	512	-	1E-6	-	-	16	1.0	3K
	AEnt (Shen, 2025)	21.7	-	75.0	37.7	33.0	–	–	✓	MATH	512	-	2E-6	7.5K	-	16	-	3K
Qwen-2.5-Math-7B	FR3E (Zheng et al., 2025b)	39.1	-	82.2	46.5	40.8	–	67.5	✗	DeepScaler, SimpleRL (3-5)	512	128	1E-6	-	-	16	-	16k
	DCPO (Yang et al., 2025b)	46.7	82.6	82.5	-	–	–	16.7	✗	MATH-DAPO-17k, MATH (3-5)	512	32	-	25k	400	-	1.0	3K
	Entropy Adv (Cheng et al., 2025)	33.7	17.6	83.1	-	-	-	pass@1: 69.8	✗	DAPO-Math-17K	512	32	1E-06	-	-	16	1.0	8K
	GRPO + D^{3}S (Wang et al., 2025a)	20.3	7.9	-	10.7	25.0	52.2	54.4	✗	DeepScaleR	-	-	5E-07	-	-	32	1.0	-
	GSPO + D^{3}S (Wang et al., 2025a)	18.3	8.3	-	11.5	28.4	54.9	53.2	✗	DeepScaleR	-	-	5E-07	-	-	32	1.0	-
	AEPO-Wang (Wang et al., 2025b)	50.0	-	-	42.4	37.5	82.0	–	✓	DAPO-17K	128	-	10-6	-	-	-	-	2K
	SimKO (Peng et al., 2025)	32.8	12.9	77.6	39.8	35.0	–	62.4	✓	MATH	1K	256	10-6	-	-	8	1.0	-
	EDGE-GRPO (Zhang et al., 2025e)	16.67	-	79.00	42.67	36.03	–	–	✓	DeepScaleR-Random-1K, DeepScaleR-Hard-1K	-	-	1E-6	2k	1000	-	-	1K
	EMPO (Zhang et al., 2025b)	20.0	-	-	37.3	40.4	78.0	65.0	✓	-	-	-	-	-	-	-	-	-
	UCAS (Xie et al., 2025)	43.3	-	85.6	48.0	37.6	–	–	✗	MATH (3-5)	512	-	1E-6	-	-	16	1.0	3K
Qwen-2.5-Math-7B-Instruct	EM_INF (Agarwal et al., 2025)	14.4	-	80.2	40.8	42.3	–	–	✗	-	-	-	-	-	-	-	0.1	-
Qwen-3 Models																		
Qwen-3-4B	PEAR (Huang et al., 2025a)	56.66	–	84.00	–	–	–	87.50	✓	GSM8K	128	-	1E-6	7,473	-	8	0.6	16K
	AER (Zhang et al., 2025d)	25.2	22.1	86.7	–	–	–	70.2	✗	open-source DeepscaleR	128	32	1E-6	-	-	-	-	8k
Qwen-3-8B	PEAR (Huang et al., 2025a)	60	–	85.40	–	–	–	92.50	✓	GSM8K	128	-	1E-6	7,473	-	8	0.6	16K
	AER (Zhang et al., 2025d)	31.4	25.1	89.4	–	–	–	75.6	✗	open-source DeepscaleR	128	32	1E-6	-	-	-	-	8k
	QAE+CLIP-Cov (Wu et al., 2025b)	46.04	37.40	–	–	–	–	90.23	✓	DAPO-Math-17K	512	32	1E-6	-	-	32	0.7	20K
	QAE+Clip-Higher (Wu et al., 2025b)	48.23	34.90	–	–	–	–	92.97	✓	DAPO-Math-17K	512	32	1E-6	-	-	32	0.7	20K
	QAE+KL-Cov (Wu et al., 2025b)	44.69	33.44	–	–	–	–	87.97	✓	DAPO-Math-17K	512	32	1E-6	-	-	32	0.7	20K

Table 2: Performance of Various Entropy-Based Methods on Mathematical Benchmarks Across Different Evaluation Metrics

Base Model	Method	AIME24	AIME25	MATH500	Olympiad	Minerva	MATH	AMC23	open-source	Train Data	Global Size	Mini Batch	Learn. Rate	Data Col.	Train. Steps	Rollout Num.	Samp. Temp.	Max Len.
DeepSeek-R1-Distill-Qwen Models																		
DeepSeek-R1-Distill-Qwen-1.5B	PEAR (Huang et al., 2025a)	acc@1: 23.33	-	acc@1: 77.20	-	-	-	acc@1: 70.00	✓	GSM8K	128	-	1E-06	7,473	-	8	0.6	16K
	CE-GPPO (Su et al., 2025a)	avg@32: 42.0	avg@32: 33.9	avg@4: 91.0	-	-	-	avg@32: 85.9	✓	KlearReasoner-MathSub-30K	-	-	1E-06	30k	-	8	-	16k
	AEnt (Shen, 2025)	avg@4: 39.2	-	avg@4: 88.2	avg@4: 59.1	avg@4: 35.9	-	-	✓	OpenR1-math	256	-	1E-06	40k	-	8	-	7K
	ASPO-Math-1.5B (Wang et al., 2025c)	avg@64: 49.0 pass@64: 80.0	avg@64: 35.1 pass@64: 70.0	avg@64: 90.5 pass@64: 94.4	avg@64: 58.8 pass@64: 66.9	avg@64: 35.1 pass@64: 50.4	-	avg@64: 87.2 pass@64: 95.0	✓	-	-	-	-	-	-	-	-	-
	Archer-Math-1.5B (Wang et al., 2025d)	avg@64: 48.7 pass@64: 83.3	avg@64: 33.8 pass@64: 70.0	avg@4: 90.8 pass@4: 94.4	avg@4: 59.3 pass@4: 67.1	avg@8: 35.7 pass@8: 51.1	-	avg@64: 86.0 pass@64: 97.5	✓	DeepScaleR, SkyworkOR1, DAPO-Math-17K	64	32	1E-06	51,800	520	16	1.0	32K
	DA-DLER-R1-1.5B (Liu et al., 2025e)	pass@1: 34.37	-	-	pass@1: 48.70	pass@1: 44.89	pass@1: 86.70	-	✓	DeepScaleR-Preview-Dataset	512	64	1E-06	40K	600	16	1.0	4000
	DLER-R1-1.5B (Liu et al., 2025e)	pass@1: 34.38	-	-	pass@1: 48.31	pass@1: 43.59	pass@1: 86.95	-	✓	DeepScaleR-Preview-Dataset	512	64	1E-06	40K	450	16	1.0	4000
	ProRL (Liu et al., 2025d)	pass@1: 48.13	pass@1: 33.33	-	pass@1: 60.22	pass@1: 47.98	pass@1: 91.89	-	✓	DeepScaleR Dataset	256	64	2E-06	40k	-	16	1.2	8k
	ADR (Zhang et al., 2025f)	pass@1: 36.5	pass@1: 23.3	pass@1: 81.0	-	-	-	-	✗	OpenMathReasoning DeepScaleR-Preview	-	-	-	-	-	-	-	8k/16k
DeepSeek-R1-Distill-Qwen-7B	TreeRL (Hou et al., 2025)	acc@1: 60.8	-	acc@1: 94.4	acc@1: 57.1	-	-	-	✓	MATH-train, NuminaMath	480	-	1.5E-06	-	-	30	1.2	8K
	CE-GPPO (Su et al., 2025a)	avg@32: 66.0	avg@32: 51.4	avg@4: 95.6	-	-	-	avg@32: 93.8	✓	KlearReasoner-MathSub-30K	-	-	1E-06	30k	-	8	-	16k
	DA-DLER-R1-7B (Liu et al., 2025e)	pass@1: 53.90	-	-	pass@1: 61.16	pass@1: 53.60	pass@1: 94.17	-	✓	DeepScaleR-Preview-Dataset	512	64	1E-06	40K	600	16	1.0	4000
	DLER-R1-7B (Liu et al., 2025e)	pass@1: 55.62	-	-	pass@1: 60.48	pass@1: 53.88	pass@1: 94.21	-	✓	DeepScaleR-Preview-Dataset	512	64	1E-06	40K	450	16	1.0	4000
LLaMA Models																		
LLaMA-3.1-8B	MENTOR (Qin et al., 2025)	avg@32: 1.2	avg@32: 0.6	-	pass@1: 8.9	pass@1: 16.2	pass@1: 30.2	-	✓	OpenR1-MATH	128	64	1E-06	220K	120	8	1.0	8k
	TTRL (Liu et al., 2025a)	pass@1: 10.0	-	pass@1: 63.7	-	-	-	-	✗	AIME24, AMC, MATH-500	-	-	-	-	-	32	0.6	3K
	ETMR (Liu et al., 2025a)	pass@1: 16.9	-	pass@1: 59.5	-	-	-	-	✗	AIME24, AMC, MATH-500	-	-	-	-	-	32	0.6	3K
LLaMA-3.1-8B-Instruct	AEPO-Dong (Dong et al., 2025a)	pass@1: 26.7	pass@1: 16.7	pass@1: 65.8	-	-	pass@1: 80.6	-	✓	-	128	16	-	-	-	16	0.6	20K
	ARPO (Dong et al., 2025b)	pass@1: 23.3	pass@1: 16.7	pass@1: 64.6	-	-	pass@1: 80.2	-	✓	Tool-Star, STILL, Tool-Star	128	16	-	-	-	16	-	4K
	EM_INF (Agarwal et al., 2025)	pass@1: 3.3	-	pass@1: 43.0	pass@1: 16.4	pass@1: 22.8	-	-	✓	-	-	-	-	-	-	-	0.1	-
	GRPO + D^2S (Wang et al., 2025a)	pass@1: 5.3	pass@1: 0.1	-	pass@1: 3.3	pass@1: 22.5	pass@1: 35.9	pass@1: 20.3	✗	DeepScaleR	-	-	5E-07	-	-	32	1.0	-
LLaMA-3.2-3B-Instruct	SimKO (Peng et al., 2025)	pass@1: 13.8	pass@1: 1.0	pass@1: 54.6	pass@1: 21.0	pass@1: 18.5	-	pass@1: 35.2	✓	GSM8K, MATH	1K	256	1E-06	-	-	8	1.0	-
	M2PO (Zheng et al., 2025a)	avg@16: 10.4	avg@16: 4.4	avg@4: 52.0	avg@4: 18.1	avg@4: 21.2	-	avg@4: 33.8	✓	DeepScaleR Math	256	512	1E-06	-	1000	8	1.0	16K
Qwen-2 Models																		
Qwen-2.5-3B	TTRL (Liu et al., 2025a)	pass@1: 7.9	-	pass@1: 72.2	-	-	-	-	✗	AIME24, AMC, MATH-500	-	-	-	-	-	64/32	0.6	3K
	MENTOR (Qin et al., 2025)	avg@32: 8.3	avg@32: 3.8	-	pass@1: 35.2	pass@1: 26.5	pass@1: 69.8	-	✓	MATH (3-5)	128	64	1E-06	-	120	8	1.0	8k
	DCPO (Yang et al., 2025b)	avg@1: 3.3	avg@1: 62.5	avg@1: 71.2	-	-	-	avg@1: 10	✓	DAPO-Math-17k, MATH (3-5)	512	32	-	25k	400	-	1.0	3K
	ETMR (Liu et al., 2025a)	pass@1: 9.2	-	pass@1: 71.7	-	-	-	-	✗	AIME24, AMC, MATH-500	-	-	-	-	-	-	-	-
Qwen-2.5-7B	GRPO+POSITION (Deng et al., 2025)	pass@4: 29.75	pass@4: 25.36	pass@4: 90	-	-	-	pass@4: 80	✗	STILL-3	512	32	1E-06	90K	-	-	1.0	8K
	GRPO+PPL (Deng et al., 2025)	pass@4: 35.48	pass@4: 24.35	pass@4: 92.4	-	-	-	pass@4: 85	✗	STILL-3	512	32	1E-06	90K	-	-	1.0	8K
	SimKO (Peng et al., 2025)	pass@1: 16.3	pass@1: 9.4	pass@1: 76.7	pass@1: 38.7	pass@1: 35.2	-	pass@1: 57.3	✓	MATH	1K	256	1E-06	-	-	8	1.0	-
	MENTOR (Qin et al., 2025)	avg@32: 18.3	avg@32: 16.5	-	pass@1: 45.2	pass@1: 34.9	pass@1: 81.4	-	✓	MATH (3-5)	128	64	1E-06	-	120	8	1.0	8k
	Clip-Cov (Cui et al., 2025)	avg@32: 22.1	avg@32: 15.8	avg@32: 80.4	avg@32: 44.1	avg@32: 41.1	-	-	✓	Eurus-2-RL-Math	256	128	5E-07	-	-	8	1.0	8K
	KL-Cov (Cui et al., 2025)	avg@32: 22.6	avg@32: 12.9	avg@32: 80.8	avg@32: 42.6	avg@32: 38.2	-	-	✓	Eurus-2-RL-Math	256	128	5E-07	-	-	8	1.0	8K
	FlowRL (Zhu et al., 2025)	avg@16: 15.41	avg@16: 10.83	avg@16: 66.96	avg@16: 34.61	avg@16: 31.41	-	-	✓	DAPO-Math-17K	512	-	1E-06	-	-	8	-	8K
	EM 1-shot (Gao et al., 2025)	-	-	avg@8: 67.4	avg@8: 33.6	avg@8: 22.1	-	avg@8: 45.6	✓	NuminaMath	64	2	2E-05	1.0	10	-	0.5	-
	FR3E (Zheng et al., 2025b)	acc: 25.2	-	acc: 79	acc: 42.1	acc: 39	-	acc: 67.5	✓	DeepScaler, SimpleRL (3-5)	512	128	1E-06	-	-	16	-	16k
Qwen-2.5-7B-Instruct	AEPO-Dong (Dong et al., 2025a)	pass@1: 33.3	pass@1: 33	pass@1: 80.4	-	-	pass@1: 90	-	✓	-	128	16	-	1K	-	16	0.6	20K
	ARPO (Dong et al., 2025b)	pass@1: 30	pass@1: 30	pass@1: 78.8	-	-	pass@1: 88.8	-	✓	STILL, Tool-Star	128	-	-	-	-	16	-	4K
	EM_INF (Agarwal et al., 2025)	pass@1: 11.1	-	pass@1: 73.8	pass@1: 38.2	pass@1: 41.2	-	-	✓	-	-	-	-	-	-	-	0.1	-
Qwen-2.5-14B	TreeRL (Hou et al., 2025)	acc: 20.8	-	acc: 81.7	acc: 44.6	-	-	-	✓	MATH-train, NuminaMath	480	-	1.5E-06	-	-	30	1.2	8K
	DCPO (Yang et al., 2025b)	avg@1: 20	avg@1: 85	avg@1: 84.6	-	-	-	avg@1: 23.3	✓	DAPO-Math-17k, MATH (3-5)	512	32	-	25k	400	-	1.0	3K
	STEER (Hao et al., 2025)	avg@32: 19.3	avg@32: 14.0	avg@1: 81.6	avg@1: 46.3	avg@1: 39.1	-	avg@32: 70.3	✓	DAPO-Math-17k	512	32	1E-06	-	150	8	1.0	3K
Qwen-2.5-32B	DAPO (Yu et al., 2025)	avg@32: 50	-	-	-	-	-	-	✓	DAPO-Math-17K	-	512	1E-06	17K	-	16	-	20K
	M2PO (Zheng et al., 2025a)	avg@16: 24.8	avg@16: 19.4	avg@4: 85.7	avg@4: 51.7	avg@4: 41.5	-	avg@4: 76.3	✓	DeepScaleR Math	256	512	1E-06	-	1000	8	1.0	16K
	Clip-Cov (Cui et al., 2025)	avg@32: 32.3	avg@32: 22.7	avg@32: 87	avg@32: 57.2	avg@32: 46	-	-	✓	Eurus-2-RL-Math	256	128	5E-07	-	-	8	1.0	8K
	KL-Cov (Cui et al., 2025)	avg@32: 36.8	avg@32: 30.8	avg@32: 84.6	avg@32: 49	avg@32: 46.3	-	-	✓	Eurus-2-RL-Math	256	128	5E-07	-	-	8	1.0	8K
	FlowRL (Zhu et al., 2025)	avg@16: 23.95	avg@16: 21.87	avg@16: 80.75	avg@16: 51.83	avg@16: 38.21	-	-	✓	DAPO	512	-	1E-06	-	-	8	-	8K
	FR3E (Zheng et al., 2025b)	acc: 40.2	-	acc: 87.4	acc: 51.7	acc: 45.6	-	acc: 80	✓	DeepScaler, SimpleRL (3-5)	512	128	1E-06	-	-	16	-	16k
Qwen-2-Math Models																		
Qwen-2.5-Math-1.5B	GRPO + D^2S (Wang et al., 2025a)	pass@1: 11.2	pass@1: 6.9	-	pass@1: 10.6	pass@1: 23.5	pass@1: 53.7	pass@1: 48.6	✗	DeepScaleR	-	-	5E-07	-	-	32	1.0	-
	GSPO + D^2S (Wang et al., 2025a)	pass@1: 11.4	pass@1: 8.2	-	pass@1: 10.5	pass@1: 22.9	pass@1: 54.0	pass@1: 48.4	✗	DeepScaleR	-	-	5E-07	-	-	32	1.0	-
	TTRL (Liu et al., 2025a)	pass@1: 15.8	-	pass@1: 73.0	-	-	-	-	✗	AIME24, AMC, MATH-500	-	-	5E-07	-	-	64/32	0.6	3K
	EDGE-GRPO (Zhang et al., 2025e)	pass@1: 10.0	-	pass@1: 73.20	pass@1: 37.33	pass@1: 29.04	-	-	✓	DeepScaleR-Random-1K, DeepScaleR-Hard-1K	-	-	1E-06	2k	1000	-	-	1K
	ETMR (Liu et al., 2025a)	pass@1: 21.0	-	pass@1: 76.9	-	-	-	-	✗	AIME24, AMC, MATH-500	-	-	-	-	-	-	-	-
	EMPO (Zhang et al., 2025b)	pass@1: 13.3	-	-	pass@1: 36.6	pass@1: 32.4	pass@1: 73.0	pass@1: 55.0	✓	-	-	-	-	-	-	-	-	-
	STEER (Hao et al., 2025)	avg@32: 17.2	avg@32: 9.7	avg@1: 75.4	avg@1: 36.9	avg@1: 28.0	-	avg@32: 61.3	✓	DAPO-Math-17k	512	32	1E-06	-	150	8	1.0	3K
	UCAS (Xie et al., 2025)	pass@1: 23.3	-	pass@1: 80.6	pass@1: 42.1	pass@1: 31.6	-	-	✓	MATH (3-5)	512	-	1E-06	-	-	16	1.0	3K
	AEnt (Shen, 2025)	acc: 21.7	-	acc: 75.0	acc: 37.7	acc: 33.0	-	-	✓	MATH	512	-	2E-06	7.5K	-	16	-	3K
Qwen-2.5-Math-7B	FR3E (Zheng et al., 2025b)	acc: 39.1	-	acc: 82.2	acc: 46.5	acc: 40.8	-	acc: 67.5	✓	DeepScaler, SimpleRL (3-5)	512	128	1E-06	-	-	16	-	16k
	ACPO (Yin et al., 2025)	acc@8: 34.2	acc@8: 16.25	acc@8: 83.4	-	-	-	acc@8: 71.9	✗	DAPO-17k	192	-	1E-06	-	-	-	-	-
	DCPO (Yang et al., 2025b)	avg@1: 46.7	avg@1: 82.6	avg@1: 82.5	-	-	-	avg@1: 16.7	✓	Math-DAPO-17k, MATH (3-5)	512	32	-	25k	400	-	1.0	3K
	CANON-Inter (Chen et al., 2025a)	avg@10: 32.7	avg@10: 18.7	pass@1: 87.0	pass@1: 51.7	-	-	-	✓	OpenR1-Math-220k (45k subset)	-	-	1E-06	45k	-	16	-	8K
	Entropy Adv (Cheng et al., 2025)	pass@1: 33.7	pass@1: 17.6	pass@1: 83.1	-	-	-	pass@1: 69.8	✗	DAPO-Math-17K	512	32	1E-06	-	-	16	1.0	8K
	GRPO + D^2S (Wang et al., 2025a)	pass@1: 20.3	pass@1: 7.9	-	pass@1: 10.7	pass@1: 25.0	pass@1: 52.2	pass@1: 54.4	✗	DeepScaleR	-	-	5E-07	-	-	32	1.0	-
	GSPO + D^2S (Wang et al., 2025a)	pass@1: 18.3	pass@1: 8.3	-	pass@1: 11.5	pass@1: 28.4	pass@1: 54.9	pass@1: 53.2	✗	DeepScaleR	-	-	5E-07	-	-	32	1.0	-
	AEPO-Wang (Wang et al., 2025b)	pass@1: 50.0	-	-	pass@1: 42.4	pass@1: 37.5	pass@1: 82.0	-	✓	DAPO-17K	128	-	1E-06	-	-	-	-	2K
	SimKO (Peng et al., 2025)	pass@1: 32.8	pass@1: 12.9	pass@1: 77.6	pass@1: 39.8	pass@1: 35.0	-	pass@1: 62.4	✓	MATH	1K	256	1E-06	-	-	8	1.0	-
	EDGE-GRPO (Zhang et al., 2025e)	pass@1: 16.67	-	pass@1: 79.00	pass@1: 42.67	pass@1: 36.03	-	-	✓	DeepScaleR-Random-1K, DeepScaleR-Hard-1K	-	-	1E-06	2k	1000	-	-	1K
	EMPO (Zhang et al., 2025b)	pass@1: 20.0	-	-	pass@1: 37.3	pass@1: 40.4	pass@1: 78.0	pass@1: 65.0	✓	-	-	-	-	-	-	-	-	-
	GRPO+POSITION (Deng et al., 2025)	pass@4: 55.33	pass@4: 30.85	pass@4: 93.20	-	-	-	pass@4: 85.00	✗	STILL-3	512	32	1E-06	90K	-	-	1.0	8K
	GRPO+PPL (Deng et al., 2025)	pass@4: 56.89	pass@4: 33.18	pass@4: 94.40	-	-	-	pass@4: 87.50	✗	STILL-3	512	32	1E-06	90K	-	-	1.0	8K
	M2PO (Zheng et al., 2025a)	avg@16: 33.3	avg@16: 17.5	avg@4: 84.0	avg@4: 47.1	avg@4: 38.1	-	avg@4: 63.8	✓	DeepScaleR	256	512	1E-06	-	1000	8	1.0	16K
	CURE (Li et al., 2025b)	avg@32: 35.5 avg@1: 18.5	-	avg@32: 83.4 avg@1: 50.5	avg@1: 48.2	-	-	avg@32: 89.7	✓	DAPO-Math-17K	512	32	1E-06	17K	110	-	1.0	2596
	STEER (Hao et al., 2025)	avg@32: 36.9	avg@32: 16.2	avg@1: 82.4	avg@1: 43.3	avg@1: 41.7	-	avg@32: 72.2	✓	DAPO-Math-17k	512	32	1E-06	-	150	8	1.0	3K
	EM 1-shot (Gao et al., 2025)	-	-	avg@8: 78.8	avg@8: 39.7	avg@8: 35.3	-	avg@8: 70.3	✓	NuminaMath	64	2	2E-05	1.0	10	-	0.5	-
	UCAS (Xie et al., 2025)	pass@1: 43.3	-	pass@1: 85.6	pass@1: 48.0	pass@1: 37.6	-	-	✓	MATH (3-5)	512	-	1E-06	-	-	16	1.0	3K
Qwen-2.5-Math-7B-Instruct	EM_INF (Agarwal et al., 2025)	pass@1: 14.4	-	pass@1: 80.2	pass@1: 40.8	pass@1: 42.3	-	-	✓	-	-	-	-	-	-	-	0.1	-
Qwen-3 Models																		
Qwen-3-1.7B	RL-ZVP (Le et al., 2025)	acc@8: 12.50 pass@8: 23.99	acc@8: 6.25 pass@8: 13.23	acc@8: 70.98 pass@8: 87.35	acc@8: 35.11 pass@8: 54.84	acc@8: 29.96 pass@8: 44.76	-	acc@8: 48.75 pass@8: 77.19	✗	MATH, NuminaMath-CoT	512	32	3E-06	27.5k	672	8	1.0	4K
	M2PO (Zheng et al., 2025a)	avg@16: 14.0	avg@16: 6.5	avg@4: 71.8	avg@4: 35.6	avg@4: 29.4	-	avg@4: 48.1	✓	DeepScaleR Math	256	512	1E-06	-	1000	8	1.0	4K
Qwen-3-30B-A3B	UloRL-A3B-128k-Yarn (Du et al., 2025)	-	avg@32: 85.1	-	-	-	-	-	✓	-	-	1K	1E-06	-	-	-	0.85	140k
Qwen-3-4B	PEAR (Huang et al., 2025a)	acc@1: 56.66	-	acc@1: 84.00	-	-	-	acc@1: 87.50	✓	GSM8K	128	-	1E-06	7,473	-	8	0.6	16K
	AER (Zhang et al., 2025d)	pass@1: 25.2	pass@1: 22.1	pass@1: 86.7	-	-	-	pass@1: 70.2	✗	DeepscaleR	128	32	1E-06	-	-	-	-	8k
	M2PO (Zheng et al., 2025a)	avg@16: 26.7	avg@16: 21.0	avg@4: 85.8	avg@4: 52.3	avg@4: 40.5	-	avg@4: 64.4	✓	DeepScaleR Math	256	512	1E-06	-	1000	8	1.0	16K
Qwen-3-8B	PEAR (Huang et al., 2025a)	acc@1: 60.00	-	acc@1: 85.40	-	-	-	acc@1: 92.50	✓	GSM8K	128	-	1E-06	7,473	-	8	0.6	16K
	AER (Zhang et al., 2025d)	pass@1: 31.4	pass@1: 25.1	pass@1: 89.4	-	-	-	pass@1: 75.6	✗	DeepscaleR	128	32	1E-06	-	-	-	-	8k
	QAE+CLIP-Cov (Wu et al., 2025b)	pass@1: 46.04	pass@1: 37.40	-	-	-	-	pass@1: 90.23	✓	DAPO-Math-17K	512	32	1E-06	-	-	32	0.7	20K
	QAE+Clip-Higher (Wu et al., 2025b)	pass@1: 48.23	pass@1: 34.90	-	-	-	-	pass@1: 92.97	✓	DAPO-Math-17K	512	32	1E-06	-	-	32	0.7	20K
	QAE+KL-Cov (Wu et al., 2025b)	pass@1: 44.69	pass@1: 33.44	-	-	-	-	pass@1: 87.97	✓	DAPO-Math-17K	512	32	1E-06	-	-	32	0.7	20K
	RL-ZVP (Le et al., 2025)	acc@8: 33.33 pass@8: 54.66	acc@8: 24.58 pass@8: 39.36	acc@8: 89.73 pass@8: 96.13	acc@8: 58.20 pass@8: 72.28	acc@8: 45.96 pass@8: 58.83	-	acc@8: 78.75 pass@8: 94.18	✗	DAPO-Math-17k, NuminaMath-CoT	512	32	3E-06	37k	1056	16	1.0	8K
	M2PO (Zheng et al., 2025a)	avg@16: 30.2	avg@16: 23.1	avg@4: 87.2	avg@4: 54.8	avg@4: 42.6	-	avg@4: 71.3	✓	DeepScaleR Math	256	512	1E-06	-	1000	8	1.0	16K