Token Hidden Reward: Steering Exploration-Exploitation in GRPO Training

Anonymous Authors[1]

Abstract

Reinforcement learning (RL) has substantially advanced the reasoning capabilities of large language models (LLMs), yet how to explicitly guide training toward exploration or exploitation remains underexplored. In this work, we start from the assumption that response confidence—the model's likelihood assigned to correct responses—is a meaningful objective for reasoning tasks. To better understand and control learning under this objective, we analyze token-level dynamics in GRPO training and introduce Token Hidden Reward (THR), a novel metric that quantifies the contribution of individual tokens to response confidence. Based on THR, we propose a THR-guided reweighting strategy that modulates the learning signal to explicitly favor either high-confidence outputs (*i.e.*, exploitation) or broader output diversity (*i.e.*, exploration). Empirically, we find that increasing confidence mostly aligns with improved greedy decoding performance (exploitation), while encouraging lower-confidence increasing consistently boosts Pass@K performance (exploration).

1. Introduction

The integration of reinforcement learning (RL) has significantly advanced the reasoning capabilities of large language models (LLMs) (Guo et al., 2025; Jaech et al., 2024; Team et al., 2023). Group Relative Policy Optimization (GRPO) (Shao et al., 2024) has emerged as a widely adopted and empirically successful method for training LLMs on complex reasoning tasks. Models like DeepSeek-R1 (Guo et al., 2025), DeepSeek-Math (Shao et al., 2024), Med-R1 (Lai et al., 2025), and Search-R1 (Jin et al., 2025) have leveraged GRPO to achieve state-of-the-art performance across diverse domains.

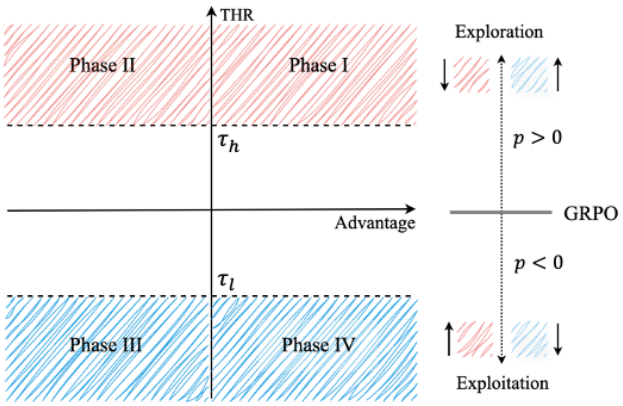

Figure 1. Tokens with positive THR tend to increase the confidence of a correct response, thus encouraging exploitation, while tokens with negative THR will reduce the confidence, thus allowing further exploration. By adjusting the weights on the tokens, we can control the exploitation-exploration tradeoff of LLM training.

Despite these successes, a central and persistent challenge in RL-driven LLM training is managing the inherent exploration-exploitation trade-off (Tang et al., 2024; Harris & Slivkins, 2025). Exploration—sampling uncertain actions to acquire novel information—is crucial for tasks demanding creativity (Lu et al., 2024) and enabling generalization to unseen test cases via scaling algorithms (Snell et al., 2024). Conversely, exploitation prioritizes optimal decision-making based on current knowledge, a preference in applications requiring high-confidence, low-variance responses, such as medical diagnosis (Wu et al., 2025). However, effectively shifting the training objective between exploration and exploitation remains an underexplored challenge.

While (Chow et al., 2024) explore the exploration-exploitation trade-off through a best-of-n training objective, their method depends on an external verifier to select the best candidate among n generations. More recently, (Deng et al., 2025) analyzed the learning dynamics of GRPO, highlighting how training impacts the confidence of correct responses. By downweighting penalties on tokens that reduce this confidence, their method improved performance under greedy decoding and better exploited model capabilities. However, their focus was primarily on negative gradients of GRPO and it's impact on exploitation, leaving other aspects underexplored.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Motivated by (Deng et al., 2025), we examine the intrinsic contribution of each token to the confidence of correct responses and explore its connection to the exploration–exploitation trade-off. We introduce Token Hidden Reward (THR), a metric that quantifies how individual tokens influence the likelihood of correct responses within the GRPO framework. By analyzing the magnitude of THR, we find that a small subset of tokens carries disproportionately high absolute THR values, while most have negligible impact. By studying the sign of THR, we propose a token reweighting strategy that adjusts learning signals based on THR. This allows the model to either reinforce observed correct responses' confidence or maintain probability mass for alternative responses (Ren & Sutherland, 2025), showing a correlation with the trade-off between exploitation and exploration. Specifically, amplifying tokens with positive THR can sometimes enhance exploitation by improving greedy decoding accuracy, whereas emphasizing tokens with negative THR encourages exploration and improves pass@K performance. Our main contributions are threefold:

- We introduce Token Hidden Reward (THR) and conduct a thorough analysis, uncovering that a small subset of tokens disproportionately influences training and that the sign of THR correlates with the exploration-exploitation trade-off.
- We propose a THR-guided reweighting strategy that effectively directs the fine-tuning process, enabling targeted emphasis on either exploitation or exploration.
- Empirical evaluations on math benchmarks confirm that THR-guided reweighting effectively guides the fine-tuning process, resulting in the successful realization of desired performance improvements.

2. Related Work

Reinforcement Learning for LLM Reasoning. Recent works have explored the use of model-generated solutions as a form of bootstrapping to strengthen the reasoning capabilities of large language models (LLMs)(Jaech et al., 2024; Guo et al., 2025; Team et al., 2025). These methods typically generate candidate solutions using a pre-trained model, then filter them based on intermediate correctness signals(Setlur et al., 2024) or final answer correctness (Guo et al., 2025; Team et al., 2025), producing high-quality data to train a new model. Building on the success of reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022), follow-up works such as GRPO (Shao et al., 2024; Guo et al., 2025) use online training to further enhance reasoning. Moreover, reinforcement learning directly incorporates the model's incorrect outputs into training, which has been found to further boost reasoning performance (Seed et al., 2025). Despite these advances, the role of model-generated outputs during training remains underexplored.

Optimizing for inference time objectives. Recent fine-tuning methods have been developed to align with various test-time objectives. For instance, some approaches treat inference-time computation as an optional post-hoc design choice (Snell et al., 2024), while others, such as (Huang et al., 2025), aim to improve best-of-n performance during training. However, the latter relies on an external verifier to select the best output among n candidates, making the implementation hard. Meanwhile, (Deng et al., 2025) focus on improving the model's most confident predictions by reducing penalties on tokens that contribute positively to correct responses, thereby enhancing greedy decoding performance. Nonetheless, their approach does not address scenarios where exploration (Dou et al., 2025) capabilities are crucial.

3. Preliminary

3.1. Notations

We use $\boldsymbol{W}$, $\boldsymbol{w}_z$, and $\boldsymbol{h}_{\boldsymbol{z}}$ to denote the token unembedding matrix, unembedding of a token $z \in \mathcal{V}$, and hidden embedding of $\boldsymbol{z} \in \mathcal{V}^*$, respectively. We let $\boldsymbol{z}_k$ be the k-th token in $\boldsymbol{z}$ and $\boldsymbol{z}_{<k}$ be the first $k-1$ tokens in $\boldsymbol{z}$. For a question $\boldsymbol{x}$, the old policy $\pi_{\theta_{\text{old}}}$ generates a group of G positive/negative samples resulting in $(\boldsymbol{x}, \{\boldsymbol{y}_i^+\}_{N^+}, \{\boldsymbol{y}_j^-\}_{N^-})$, where $N^+ + N^- = G$. Lastly, we denote by $\boldsymbol{e}_z \in \mathbb{R}^{|\mathcal{V}|}$ the standard basis vector corresponding to $z \in \mathcal{V}$.

3.2. GRPO

GRPO loss, introduced in DeepSeek-Math (Guo et al., 2025) and DeepSeek-R1 (Shao et al., 2024), enhances fine-tuning by refining how reward and loss are calculated. Concretely, unlike traditional Proximal Policy Optimization (PPO) (Schulman et al., 2017), GRPO eliminates the need for value function estimation, employing group-relative rewards for a more nuanced optimization process.

For a query-answer pair $(\boldsymbol{x}, \boldsymbol{a})$, the policy π_θ samples G responses $\{\boldsymbol{y}_i\}_{i=1}^G$. Each $\boldsymbol{y}_i$ consists of a sequence of $|\boldsymbol{y}_i|$ tokens, and we denote $\boldsymbol{y}_{i,<k}$ the subsequence of the first $k-1$ tokens. Let r_i denote the reward for response $\boldsymbol{y}_i$. The advantage of the i-th response is computed by normalizing the group-level rewards $\{r_i\}_{i=1}^G$ and is the same for each token $k = 1, \ldots, |\boldsymbol{y}_i|$. Concretely, $\hat{A}_{i,k} := \frac{r_i - \mu}{\sigma}$, with $\mu = \widehat{\mathbb{E}}[\{r_i\}_{i=1}^G]$ and $\sigma = \sqrt{\widehat{\text{Var}}[\{r_i\}_{i=1}^G]}$ being the empirical average and standard deviation of the rewards. The GRPO

objective $\mathcal{J}_{\text{GRPO}}(\theta)$ is then defined as:

$$\mathop{\mathbb{E}}_{\substack{(\boldsymbol{x},\boldsymbol{a})\sim\mathcal{D} \\ \{\boldsymbol{y}_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|\boldsymbol{x})}} \Bigg[\frac{1}{\sum_{i=1}^{G}|\boldsymbol{y}_i|}\sum_{i=1}^{G}\sum_{k=1}^{|\boldsymbol{y}_i|}\min\left(\gamma_{i,k}(\theta)\hat{A}_{i,k},\right.$$
$$\left.\hat{A}_{i,k}\cdot\text{clip}\left(\gamma_{i,k}(\theta),1-\varepsilon,1+\varepsilon\right)\right)\Bigg] \quad (1)$$

where ε is a clipping hyperparameter, $\text{clip}(\cdot)$ is the clipping operation, and $\gamma_{i,k}(\theta)=\frac{\pi_\theta(\boldsymbol{y}_{i,k}|\boldsymbol{x},\boldsymbol{y}_{i,<k})}{\pi_{\theta_{\text{old}}}(\boldsymbol{y}_{i,k}|\boldsymbol{x},\boldsymbol{y}_{i,<k})}$ is the likelihood ratio between the current policy π_θ and the old policy $\pi_{\theta_{\text{old}}}$.

3.3. Likelihood Change of Correct Response in GRPO

A recent study (Deng et al., 2025) analyzed the learning dynamics of GRPO and examined how the likelihood of a correct response $\boldsymbol{y}_i^+$ evolves during training, leading to the formulation of the following theorem which is proved leveraging a model of unconstrained features (Yang et al., 2017):

Theorem 3.1. *For any question $\boldsymbol{x}$, at any time $t\geq 0$ of training, and any correct response $\boldsymbol{y}_i^+, i\in[N^+]$, in addition to the dependence on token unembeddings, the likelihood change $\frac{d}{dt}\ln\pi_{\theta(t)}(\boldsymbol{y}_i^+|\boldsymbol{x})$ exhibits increased laziness (that is, has smaller magnitude) as the following quantity increases:*

$$p^-\sum_{k=1}^{|\boldsymbol{y}_i^+|}\sum_{j=1}^{N^-}\sum_{k'=1}^{|\boldsymbol{y}_j^-|}\underbrace{\alpha^-_{k,k'}\cdot\langle\mathbf{h}_{\boldsymbol{x},\boldsymbol{y}^+_{i,<k}},\mathbf{h}_{\boldsymbol{x},\boldsymbol{y}^-_{j,<k'}}\rangle}_{\textit{Negative Token Hidden Reward}}-$$
$$p^+\sum_{k=1}^{|\boldsymbol{y}_i^+|}\sum_{i'=1}^{N^+}\sum_{k''=1}^{|\boldsymbol{y}_{i'}^+|}\underbrace{\alpha^+_{k,k''}\cdot\langle\mathbf{h}_{\boldsymbol{x},\boldsymbol{y}^+_{i,<k}},\mathbf{h}_{\boldsymbol{x},\boldsymbol{y}^+_{i',<k''}}\rangle}_{\textit{Positive Token Hidden Reward}}. \quad (2)$$

where

$$\alpha^+_{k,k''}(t)=$$
$$\left\langle\mathbf{e}_{\boldsymbol{y}^+_{i,k}}-\pi_{\theta(t)}(\cdot\mid\boldsymbol{x},\boldsymbol{y}^+_{i,<k}),\mathbf{e}_{\boldsymbol{y}^+_{i',k''}}-\pi_{\theta(t)}(\cdot\mid\boldsymbol{x},\boldsymbol{y}^+_{i',<k''})\right\rangle,$$
$$\alpha^-_{k,k'}(t)=$$
$$\left\langle\mathbf{e}_{\boldsymbol{y}^+_{i,k}}-\pi_{\theta(t)}(\cdot\mid\boldsymbol{x},\boldsymbol{y}^+_{i,<k}),\mathbf{e}_{\boldsymbol{y}^-_{j,k'}}-\pi_{\theta(t)}(\cdot\mid\boldsymbol{x},\boldsymbol{y}^-_{j,<k'})\right\rangle.$$

which quantify the similarity of token-level prediction error across responses.

This theorem offers a theoretical framework for understanding how tokens in reinforcement learning (GRPO in this case) generated responses affect the likelihood of a correct response.

4. Token Hidden Reward

In this paper, we adopt the setting of (Deng et al., 2025) to analyze the dynamics of the log-likelihood of a correct response, $\frac{d}{dt}\ln\pi_{\theta(t)}(\boldsymbol{y}_i^+\mid\boldsymbol{x})$, reflecting the objective of RL training to increase the probability of producing accurate outputs. Building on Theorem 3.1 from their work, we introduce the concept of *token hidden reward (THR)* to more precisely capture the impact of individual tokens on this likelihood.

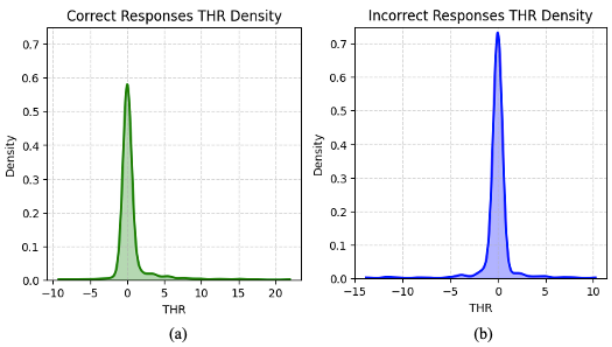

Figure 2. Density of THR scores for Qwen2.5-Math-1.5B. For both correct responses (a) and incorrect responses (b), we observe that only a small subset of tokens exhibits significantly high THR values. Notably, both types of responses contain tokens with both positive and negative THR scores.

Definition 4.1. Given a question $\boldsymbol{x}$ and a correct response $\boldsymbol{y}_i^+$, for any token $\boldsymbol{y}_{k'}$ in another response $\boldsymbol{y}$, the *THR* quantifies that token's contribution to the change in the likelihood of the correct response $\frac{d}{dt}\ln\pi_{\theta(t)}(\boldsymbol{y}_i^+\mid\boldsymbol{x})$. Formally, the hidden reward for the k'-th token is defined as:

$$\text{THR}(\boldsymbol{y}_i^+,\boldsymbol{y},k')=$$
$$\sum_{k=1}^{|\boldsymbol{y}_i^+|}\begin{cases}\alpha_{k,k'}\cdot\langle\mathbf{h}_{\mathbf{x},\boldsymbol{y}^+_{i,<k}},\mathbf{h}_{\mathbf{x},\boldsymbol{y}_{<k'}}\rangle, & \text{if } r(\boldsymbol{y})=1,\\ -\alpha_{k,k'}\cdot\langle\mathbf{h}_{\mathbf{x},\boldsymbol{y}^+_{i,<k}},\mathbf{h}_{\mathbf{x},\boldsymbol{y}_{<k'}}\rangle, & \text{otherwise.}\end{cases} \quad (3)$$

In view of Theorem 3.1, a larger THR corresponds to a greater increase in likelihood. The negative sign for incorrect responses ($r(\boldsymbol{y})=0$) reflects the fact that reinforcement learning penalizes those responses. In the GRPO setting, we extend the definition of token hidden reward to this group context as follows:

Corollary 4.2. *Given a question $\boldsymbol{x}$ and the set of correct responses $\{\boldsymbol{y}_i^+\}_{N^+}$, for any token $\boldsymbol{y}_{k'}$ in a response $\boldsymbol{y}_j$ (where $\boldsymbol{y}_j\in\{\boldsymbol{y}_i^+\}_{N^+}\cup\{\boldsymbol{y}_j^-\}_{N^-}$), the token hidden reward is defined as its contribution to the change in the likelihood of the group of correct responses $\sum_{i=1}^{N^+}\frac{1}{|\boldsymbol{y}_i^+|}\frac{d}{dt}\ln\pi_{\theta(t)}(\boldsymbol{y}_i^+\mid\boldsymbol{x})$. Formally, the k'-th token's contribution to the change of likelihood of the group of correct responses is:*

$$\text{THR}(\boldsymbol{y}_j,k')=\sum_{i=1}^{N^+}\frac{1}{|\boldsymbol{y}_i^+|}\text{THR}(\boldsymbol{y}_i^+,\boldsymbol{y}_j,k').$$

In Corollary 4.2, the sign of $\text{THR}(\boldsymbol{y}_j,k')$ indicates whether the token positively or negatively contribute to the likeli-

hood of the correct response. Specifically, as highlighted in red shading in Figure 1, increasing the influence (e.g., by reweighting the token's advantage $w \cdot \hat{A}_{k'}, w > 1$) of tokens with positive THR reduces the value of the quantity in Equation (2), thereby boosting the likelihood of correct responses and leading to better exploitation. Conversely, as shown in blue shading in Figure 1, increasing the influence of tokens with negative THR increases the value of the quantity in Equation (2), reducing the likelihood of correct responses and encouraging exploration. Thus, the magnitude of $\mathrm{THR}(\boldsymbol{y}_{k'})$ reflects the strength of each token's influence on the likelihood.

5. THR Analysis

In this section, we analyze the Token Hidden Reward (THR). We present the density of token THR scores in Figure 2.

Dominant Tokens. In both correct and incorrect responses, the majority of tokens have THR scores clustered around zero. However, a small subset of tokens exhibit significantly larger THR values, indicating that these tokens dominate the training dynamics.

Sign of THR. As shown in Figure 2, both correct responses (a) and incorrect responses (b) contain tokens with both positive and negative THR scores, indicating that tokens in either type of response can either increase or decrease the confidence in the correct response.

5.1. Impact of Dominant Tokens.

In this section, we study the impact of dominant tokens. First we define the dominant tokens as tokens with THR score higher than a threshold $\mathrm{THR} > \tau$. We detail the selection of τ in Section 6.1.

Dominant Token Training We first only use these influencial tokens by setting other tokens' advantage to zero.

$$\mathop{\mathbb{E}}_{\substack{(\boldsymbol{x},\boldsymbol{a})\sim\mathcal{D} \\ \{\boldsymbol{y}_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|\boldsymbol{x})}} \Bigg[\frac{1}{\sum_{i=1}^{G}|\boldsymbol{y}_i|}\sum_{i=1}^{G}\sum_{k=1}^{|\boldsymbol{y}_i|}\mathbb{1}[|\mathrm{THR}_{i,k}|>\tau]\cdot \min\Big(\gamma_{i,k}(\theta)\hat{A}_{i,k}, \hat{A}_{i,k}\cdot \mathrm{clip}\left(\gamma_{i,k}(\theta), 1-\varepsilon, 1+\varepsilon\right)\Big)\Bigg] \quad (4)$$

we report the results of training with Equation (4) in Table 1, where THR indicate only train with the influencial tokens, where the performance is similar to that of GRPO which training with all tokens and achieve the same performance on average, indicating these influential tokens dominate the RL training.

Relationship with Entropy. Since a confident (low-entropy) token will have a small $\mathbf{e}_{\boldsymbol{y}_{k'}} - \pi(\cdot|\boldsymbol{x}, \boldsymbol{y}_{<k'})$, thus the resulting α in Definition 4.1 tends to be close to zero, leading to a low THR. We analyze the overlap between tokens with high THR scores and those with high entropy. Specifically, for each sample, we select the same number of high-entropy tokens as high-THR tokens, compute their overlap rate, and plot the kernel density estimate (Chen, 2017) of the resulting overlap scores in Figure 3. The results show a consistently high overlap ratio—often around 90%—indicating a strong correlation between THR and entropy. This finding is consistent with the observation by (Wang et al., 2025), who demonstrated that training on only the top 20% of high-entropy tokens was sufficient to achieve strong performance.

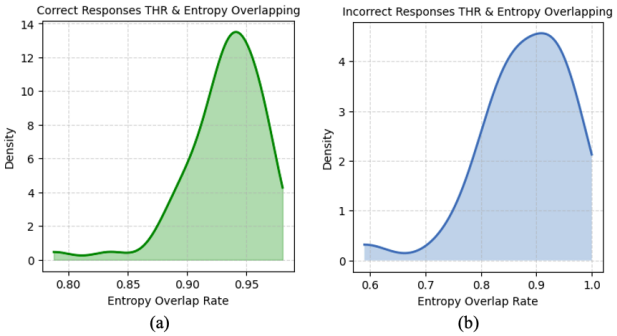

Figure 3. Overlap between high THR and high entropy tokens. For each sample, we quantify the overlap between tokens with high THR and high entropy, and plot the resulting density. The distribution shows a pronounced peak near 90%, highlighting a strong token-level association between these two metrics.

5.2. Impact of THR Signs on Exploration and Exploitation

In this section, we analyze the sign of THR and its impact on exploration and exploitation. We begin by informally defining exploration and exploitation in our context.

Exploration. A lower increase in the likelihood of correct responses preserves some probability mass for other responses, thus promoting exploration.

Exploitation. A higher increase in the likelihood of generating correct responses strengthens confidence in those observed correct answers, thereby reinforcing exploitation.

To control the exploration and exploitation, we propose a token reweighting strategy to switch between exploration and exploitation during training. As illustrated in Figure 1, to promote exploitation, we increase the weight of tokens with positive THR and decrease the weight of tokens with negative THR. In contrast, by reversing the reweighting scheme, we can encourage exploration in RL training. For influential tokens, we introduce a reweighting score highlighted in red:

$$\mathbb{1}[|\mathrm{THR}_{i,k}|>\tau]\cdot(1+\mathrm{sign}(\mathrm{THR}_{i,k})\cdot p) \quad (5)$$
$$\min\Big(\gamma_{i,k}(\theta)\hat{A}_{i,k}, \hat{A}_{i,k}\cdot \mathrm{clip}\left(\gamma_{i,k}(\theta), 1-\varepsilon, 1+\varepsilon\right)\Big)$$

As illustrated in Figure 1, setting $p > 0$ amplifies the im-

Base Model + Configuration	AIME24	AMC23	MATH500	Minerva	Olympiad	Avg.
Qwen2.5-0.5B-Ins						
Base	0.0	2.5	33.4	4.4	7.0	9.5
GRPO	0.0	7.5	33.8	8.8	**9.9**	12.0
THR	0.0	15.0	<u>34.6</u>	8.1	7.6	13.1
THR ($p = -0.2$)	0.0	**20.0**	34.0	<u>9.9</u>	<u>8.9</u>	**14.6**
THR ($p = 0.2$)	0.0	<u>17.5</u>	**35.6**	**11.0**	6.5	<u>14.1</u>
Qwen2.5-Math-1.5B						
Base	3.3	20.0	39.6	7.7	24.9	19.1
GRPO	13.3	57.5	**71.8**	29.0	34.1	41.1
Pos Only	10.0	57.5	70.6	30.1	31.0	39.8
THR	13.3	55.0	70.8	<u>32.4</u>	<u>34.1</u>	41.1
THR ($p = -0.1$)	13.3	<u>60.0</u>	70.6	32.0	32.7	<u>41.7</u>
THR ($p = 0.1$)	13.3	**62.5**	<u>71.4</u>	**33.1**	**34.5**	**43.0**

Table 1. Exploitation Results. Pass@1 accuracy (%) using greedy decoding across different methods and datasets. **Bold** indicates the best performance, while <u>underline</u> marks the second-best.

pact of tokens in Phase I and Phase II while diminishing the influence of Phase III and Phase IV, thereby promoting exploitation. In contrast, choosing $p < 0$ reverses this effect, shifting the emphasis toward exploration. We detail the results of training with this reweighting scheme in Section 6.2.

6. Experiments

6.1. Implementation Details

Models. We select models ranging from 0.5B to 1.5B parameters (Yang et al., 2024a). For the 0.5B model, we use Qwen-0.5B-Ins, as prior work (Zeng et al., 2025) suggests that small base models may struggle to follow formatted prompts. For the 1.5B model, we adopt Qwen2.5-Math-1.5B (Yang et al., 2024b). All models are fine-tuned using GRPO and THR with identical reinforcement learning hyperparameters.

Training Setup. We use the MATH dataset (levels 3–5) (Hendrycks et al., 2021) to train the model. To accelerate training, we use *dynamic sampling* (Yu et al., 2025), which discards samples with zero advantage and continues sampling new questions until a full batch is constructed. For the 0.5B model, training is conducted on two A6000 GPUs with a batch size of 32, a maximum rollout length of 2500 tokens, a learning rate of $5e^{-7}$, and a mini-batch size of 16—resulting in two iteration updates per training step. Given the increased sample efficiency introduced by dynamic sampling (looping 2–3× more questions per batch), we train for 60 steps. For the 1.5B model, we utilize four A100 GPUs with a batch size of 256, learning rate $1e^{-6}$ and a mini-batch size of 64, leading to four iteration updates per step and we train for 40 steps, which corresponds to approximately two effective epochs when considering the increased question throughput from dynamic sampling. Across all models, we generate 8 rollouts per prompt. We use a default sampling temperature of 1.0, a clipping ratio of 0.2, and set the KL loss coefficient to 1×10^{-4}. The Qwen-Math model (Yang et al., 2024b) uses its full context length of 3072 tokens for rollouts.

Selection of τ. For Qwen-Math-1.5B, we follow (Deng et al., 2025)'s Eq. (8), setting τ to the average impact of the i'-th correct response's tokens on the likelihoods of other correct responses. For Qwen-0.5B-Ins, τ is determined by selecting the top 20% of tokens based on absolute impact values, and additionally including the top 20% of tokens with the highest entropy.

Evaluation setup. Since exploitation focuses on making the best decisions based on existing knowledge (Harris & Slivkins, 2025), we assess the exploitation ability of fine-tuned models by measuring their greedy decoding accuracy on five widely used math benchmarks: AIME 2024 (Veeraboina, 2023), AMC, MATH500 (Hendrycks et al., 2021), and Minerva Math (Lewkowycz et al., 2022). To evaluate exploration, we report the unbiased Pass@K accuracy (Chen et al., 2021) using temperature 1.0 and top-p 1.0 on the more challenging AIME and AMC datasets, which require diverse solution attempts. Pass@K is defined as $\text{Pass@}K = \mathbb{E}_{x\sim D}\left[1 - \frac{\binom{M-C}{K}}{\binom{M}{K}}\right]$, where $M \geq K$ is the number of model responses per question, and C denotes the number of correct responses among them. For the greedy decoding performance of Qwen2.5-0.5B-Ins, we report the best accuracy across multiple checkpoints due to significant fluctuations during training. For all other settings, we report the performance at the final checkpoint.

Method	Qwen2.5-0.5B-Instruct Pass@K									Qwen2.5-Math-1.5B Pass@K								
	1	2	4	8	16	32	64	128	256	1	2	4	8	16	32	64	128	256
AIME 2024																		
Base	0.1	0.2	0.4	0.8	1.6	3.1	5.6	9.8	16.7	3.3	6.3	11.3	18.5	27.4	36.4	44.3	49.6	53.3
GRPO	**0.4**	**0.8**	**1.5**	**2.9**	**5.4**	**10.0**	**17.2**	**27.3**	**36.7**	11.4	17.7	24.3	30.5	36.7	43.4	50.0	56.0	63.3
Neg Only	0.2	0.5	0.9	1.8	3.3	5.9	9.7	14.9	23.3	9.9	16.0	23.1	30.2	36.7	42.8	48.1	52.9	56.7
THR	**0.4**	0.7	**1.5**	**2.9**	**5.4**	9.7	15.7	22.0	26.7	10.6	16.7	23.4	30.2	37.2	44.8	51.9	58.5	63.3
THR ($p < 0$)	**0.4**	**0.8**	**1.5**	**2.9**	**5.4**	9.4	14.9	21.5	30.0	**11.9**	**18.2**	**24.9**	**31.2**	**37.9**	**45.3**	**52.9**	**61.2**	**70.0**
THR ($p > 0$)	**0.4**	0.7	1.4	2.6	4.7	8.1	12.9	19.9	30.0	8.4	13.6	20.0	27.0	34.7	43.1	50.8	57.6	63.3
AMC23																		
Base	4.1	7.8	14.0	23.4	36.1	50.6	64.4	75.4	82.5	15.3	26.7	42.1	58.6	72.3	81.9	88.8	94.3	97.5
GRPO	11.4	18.7	28.3	39.7	52.3	64.5	74.9	81.8	85.0	46.6	59.1	70.0	78.9	85.5	90.2	93.7	96.0	97.5
Neg Only	7.7	13.7	22.6	34.4	48.4	63.2	76.6	87.5	**95.0**	44.0	56.9	68.0	76.5	83.0	88.5	92.3	94.3	95.0
THR	**12.0**	**20.2**	**30.8**	**43.0**	56.1	68.6	79.5	88.0	92.5	44.8	57.8	69.1	78.2	85.1	90.1	93.6	95.9	97.5
THR ($p < 0$)	**12.0**	20.1	30.6	42.7	**56.5**	**70.8**	**82.7**	**89.6**	92.5	**47.9**	**61.0**	**72.2**	**81.1**	**87.3**	**91.6**	**95.1**	**98.0**	**100.0**
THR ($p > 0$)	11.1	18.8	29.2	41.9	56.0	69.3	80.1	87.5	92.5	41.4	54.8	66.8	76.6	84.2	89.5	93.2	95.8	97.5
Average																		
Base	2.1	4.0	7.2	12.1	18.9	26.9	35.0	42.6	49.6	9.3	16.5	26.7	38.6	49.9	59.2	66.6	71.9	75.4
GRPO	5.9	9.8	14.9	21.3	28.9	37.2	46.1	54.6	60.9	29.0	38.4	47.2	54.7	61.1	66.8	71.9	76.0	80.4
Neg Only	4.0	7.1	11.8	18.1	25.9	34.6	43.2	51.2	59.2	27.0	36.5	45.6	53.4	59.9	65.6	70.2	73.6	75.9
THR	**6.2**	**10.5**	**16.2**	**23.0**	30.8	39.2	47.6	55.0	59.6	27.7	37.3	46.3	54.2	61.2	67.5	72.8	77.2	80.4
THR ($p < 0$)	**6.2**	10.4	16.1	22.8	**30.9**	**40.1**	**48.8**	**55.6**	**61.3**	**29.9**	**39.6**	**48.6**	**56.2**	**62.6**	**68.5**	**74.0**	**79.6**	**85.0**
THR ($p > 0$)	5.8	9.8	15.3	22.2	30.4	38.7	46.5	53.7	61.3	24.9	34.2	43.4	51.8	59.5	66.3	72.0	76.7	80.4

Table 2. Exploration Results. Pass@K results for Qwen2.5-0.5B-Instruct and Qwen2.5-Math-1.5B are reported on the AIME 2024 and AMC23 datasets, along with their average. **Bold** indicates the best performance. Specifically, $p < 0$ corresponds to $p = -0.2$ for the 0.5B model and $p = -0.1$ for the 1.5B model, while $p > 0$ corresponds to $p = 0.2$ and $p = 0.1$ respectively. Utilizing $p < 0$ leads to consistently higher Pass@K scores, indicating improved exploration capabilities.

6.2. Results

Impact of Dominant Tokens. Training exclusively with dominant tokens—those associated with high absolute THR scores—results in performance comparable to the original GRPO. As shown in Table 1, vanilla THR matches GRPO in accuracy on Qwen2.5-Math-1.5B and even surpasses it on Qwen2.5-0.5B-Instruct. Similarly, the Pass@K results in Table 2 show that vanilla THR performs on par with GRPO and notably exceeds it at higher Pass@K ($K \geq 16$) on Qwen2.5-Math-1.5B. These findings indicate that dominant tokens play a critical role in guiding the training process.

$p > 0$ **for Exploitation.** To emphasize exploitation, we set $p > 0$ to amplify the influence of tokens with positive THR values while suppressing those with negative ones. We evaluate model accuracy based on its most confident response using greedy decoding. As shown in Table 1, on Qwen2.5-Math-1.5B, THR ($p = 0.1$) achieves the best performance across all datasets, improving the average score by 1.9 % compared to the vanilla THR. This highlights the effectiveness of enhancing exploitation by reinforcing the impact of Phases I & II. Similarly, on Qwen2.5-0.5B-Ins, THR ($p = 0.2$) boosts average performance by 1% over vanilla THR. Although THR ($p = 0.2$) is outperformed by THR ($p = -0.2$) overall, it yields better in-domain results on the Math500 dataset, validating the correlation between boosting correct response confidence and achieving exploitation. Notably, while setting $p > 0$ prioritizes exploitation over exploration, it maintains comparable accuracy at higher K in Pass@K, with performance close to both vanilla THR and GRPO, as shown in Table 2.

$p < 0$ **for Exploration.** To assess exploration, we examine Pass@K at higher values of K on the challenging AIME and AMC benchmarks. A negative p increases the weight of tokens with negative THR scores, leaving more probability mass for exploration. As shown in Table 2, setting $p < 0$ consistently improves average Pass@K performance for both Qwen2.5-0.5B-Instruct and Qwen2.5-Math-1.5B models. The gains are especially pronounced on the larger Qwen2.5-Math-1.5B model at higher K values—for instance, surpassing the best baseline by 2.4% at Pass@128 and 5% at Pass@256. In almost all settings, $p < 0$ outperforms other configurations, with the sole exception of GRPO on AIME 2024 for Qwen2.5-0.5B-Instruct. Nevertheless, it still surpasses the vanilla THR, further validating its effectiveness in enhancing exploration. Notably, while setting $p < 0$ encourages exploration, it still achieves competitive greedy accuracy—outperforming both vanilla THR and GRPO, and even surpassing $p > 0$ on Qwen2.5-0.5B-Ins, as shown in Table 1. This suggests that allowing greater exploration can be beneficial.

Ablation Study on Positive and Negative-Only Training. We further investigate the impact of training with only positive or negative tokens by modifying $\hat{A}_{i,k}$. In the "Pos Only" setting, we set all values where $\hat{A}_{i,k} < 0$ to 0, thereby increasing the confidence of correct responses only. Conversely, in the "Neg Only" setting, we set all values where $\hat{A}_{i,k} > 0$ to 0, which reduces the confidence of incorrect responses without reinforcing correct ones. As shown in Table 1, "Pos Only" results in a 1.3% drop in average performance compared to GRPO, indicating that negative gradients also contribute to boosting confidence in correct responses.

As shown in Table 2, "Neg Only" underperforms in most cases. For example, on AMC23 with Qwen2.5-Math-1.5B, it achieves a Pass@256 of 56.7%, compared to 63.3% for both GRPO and vanilla THR. While "Neg Only" yields moderate improvements over the Base model on average—indicating that suppressing incorrect responses provides some exploratory value—positive tokens still play a critical role in enhancing exploration. By selectively incorporating informative tokens, THR with $p < 0$ achieves substantially better exploration performance than "Neg Only" alone.

7. Conclusion

In this work, we addressed the challenge of guiding GRPO fine-tuning of LLMs toward either exploration or exploitation by introducing THR, a metric that quantifies the impact of individual tokens on the likelihood of correct responses. Our analysis revealed that a small subset of tokens with high THR magnitudes disproportionately influences training dynamics, while the sign of THR correlates with controlling towards exploration (negative sign) or exploitation (positive sign). By proposing a THR-guided reweighting strategy, we empirically demonstrated the ability to steer training explicitly: amplifying positive THR tokens enhances exploitation, improving greedy decoding accuracy, while emphasizing negative THR tokens promotes exploration, boosting Pass@K performance. Our findings provide a principled framework for controlling the exploration-exploitation balance in RL-driven LLM training, opening avenues for more targeted fine-tuning strategies in reasoning-intensive applications. Future work could explore extending THR to other RL algorithms and more systematic ways of tuning the parameter p perhaps adaptively per iteration.

References

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Chen, Y.-C. A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1):161–187, 2017.

Chow, Y., Tennenholtz, G., Gur, I., Zhuang, V., Dai, B., Thiagarajan, S., Boutilier, C., Agarwal, R., Kumar, A., and Faust, A. Inference-aware fine-tuning for best-of-n sampling in large language models. *arXiv preprint arXiv:2412.15287*, 2024.

Deng, W., Ren, Y., Li, M., Sutherland, D. J., Li, X., and Thrampoulidis, C. On the effect of negative gradient in group relative deep reinforcement optimization. *arXiv preprint arXiv:2505.18830*, 2025.

Dou, S., Wu, M., Xu, J., Zheng, R., Gui, T., Zhang, Q., and Huang, X. Improving rl exploration for llm reasoning through retrospective replay. *arXiv preprint arXiv:2504.14363*, 2025.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Harris, K. and Slivkins, A. Should you use your large language model to explore or exploit? *arXiv preprint arXiv:2502.00225*, 2025.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.

Huang, A., Block, A., Liu, Q., Jiang, N., Krishnamurthy, A., and Foster, D. J. Is best-of-n the best of them? coverage, scaling, and optimality in inference-time alignment. *arXiv preprint arXiv:2503.21878*, 2025.

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

Jin, B., Zeng, H., Yue, Z., Yoon, J., Arik, S., Wang, D., Zamani, H., and Han, J. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Lai, Y., Zhong, J., Li, M., Zhao, S., and Yang, X. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*, 2025.

Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., et al. Solving quantitative reasoning problems with language models. *Advances in*

Neural Information Processing Systems, 35:3843–3857, 2022.

Lu, L.-C., Chen, S.-J., Pai, T.-M., Yu, C.-H., Lee, H.-y., and Sun, S.-H. Llm discussion: Enhancing the creativity of large language models via discussion framework and role-play. *arXiv preprint arXiv:2405.06373*, 2024.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Ren, Y. and Sutherland, D. J. Learning dynamics of llm finetuning. In *International Conference on Learning Representations*, 2025.

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

Seed, B., Yuan, Y., Yue, Y., Wang, M., Zuo, X., Chen, J., Yan, L., Xu, W., Zhang, C., Liu, X., et al. Seed-thinking-v1. 5: Advancing superb reasoning models with reinforcement learning. *arXiv preprint arXiv:2504.13914*, 4, 2025.

Setlur, A., Nagpal, C., Fisch, A., Geng, X., Eisenstein, J., Agarwal, R., Agarwal, A., Berant, J., and Kumar, A. Rewarding progress: Scaling automated process verifiers for llm reasoning. *arXiv preprint arXiv:2410.08146*, 2024.

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

Snell, C., Lee, J., Xu, K., and Kumar, A. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Tang, H., Hu, K., Zhou, J., Zhong, S. C., Zheng, W.-L., Si, X., and Ellis, K. Code repair with llms gives an exploration-exploitation tradeoff. *Advances in Neural Information Processing Systems*, 37:117954–117996, 2024.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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

Veeraboina, H. Aime problem set 1983-2024, 2023. URL https://www.kaggle.com/datasets/hemishveeraboina/aime-problem-set-1983-2024.

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

Wu, J., Deng, W., Li, X., Liu, S., Mi, T., Peng, Y., Xu, Z., Liu, Y., Cho, H., Choi, C.-I., et al. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*, 2025.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a.

Yang, A., Zhang, B., Hui, B., Gao, B., Yu, B., Li, C., Liu, D., Tu, J., Zhou, J., Lin, J., Lu, K., Xue, M., Lin, R., Liu, T., Ren, X., and Zhang, Z. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024b.

Yang, Z., Dai, Z., Salakhutdinov, R., and Cohen, W. W. Breaking the softmax bottleneck: A high-rank rnn language model. *arXiv preprint arXiv:1711.03953*, 2017.

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

Zeng, W., Huang, Y., Liu, Q., Liu, W., He, K., Ma, Z., and He, J. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*, 2025.