Diversity or Precision? A Deep Dive into Next Token Prediction

Haoyuan Wu[1,2], Hai Wang[1,†], Jiajia Wu[1], Jinxiang Ou[1], Keyao Wang[1],
Weile Chen[1], Zihao Zheng[1], Bei Yu[2]
[1]**LLM Department, Tencent** [2]**The Chinese University of Hong Kong**

Abstract

Recent advancements have shown that reinforcement learning (RL) can substantially improve the reasoning abilities of large language models (LLMs). The effectiveness of such RL training, however, depends critically on the exploration space defined by the pre-trained model's token-output distribution. In this paper, we revisit the standard cross-entropy loss, interpreting it as a specific instance of policy gradient optimization applied within a single-step episode. To systematically study how the pre-trained distribution shapes the exploration potential for subsequent RL, we propose a generalized pre-training objective that adapts on-policy RL principles to supervised learning. By framing next-token prediction as a stochastic decision process, we introduce a reward-shaping strategy that explicitly balances diversity and precision. Our method employs a positive reward scaling factor to control probability concentration on ground-truth tokens and a rank-aware mechanism that treats high-ranking and low-ranking negative tokens asymmetrically. This allows us to reshape the pre-trained token-output distribution and investigate how to provide a more favorable exploration space for RL, ultimately enhancing end-to-end reasoning performance. Contrary to the intuition that higher distribution entropy facilitates effective exploration, we find that imposing a precision-oriented prior yields a superior exploration space for RL.

1 Introduction

Recent advancements have demonstrated that reinforcement learning (RL) (Bai et al., 2022; Guo et al., 2025) can significantly enhance the reasoning capabilities of large language models (LLMs) (Google DeepMind, 2025; Guo et al., 2025; Anthropic, 2025; Kimi et al., 2025). By utilizing verifiable rewards, such as passing unit tests or deriving correct mathematical solutions, LLMs evolve from merely mimicking human data to actively searching for optimal reasoning paths (Guo et al., 2025). On-policy training paradigms have proven effective in unlocking the potential of pre-trained LLMs, prompting researchers to investigate how token output distributions influence RL. Recent studies (Wang et al., 2025; Zhu et al., 2025b; Cui et al., 2025; Gandhi et al., 2025) indicate that uncertainty in chain-of-thought reasoning is concentrated within a small subset of high-entropy forking tokens that govern pivotal decisions, while the majority of tokens exhibit low entropy. This observation underscores the critical impact of the pre-trained model's output distribution on subsequent RL outcomes.

Concurrently, researchers have explored next-token and next-segment reasoning objectives to derive self-supervised signals from massive unlabeled pre-training corpora (Zelikman et al., 2024; Dong et al., 2025; Li et al., 2025; Xing et al., 2025). Applying RL to the pre-training corpus suggests a theoretical bridge connecting pre-training and RL. Specifically, next-token prediction can be reformulated as a reasoning task optimized via RL algorithms, where the model receives verifiable rewards for accurately predicting the subsequent token according to a given context. Notably, if the intermediate reasoning process is omitted, resulting in the direct generation of the answer, this procedure becomes analogous to standard pre-training. From the perspective of policy optimization, next-token prediction serves a foundational role by defining the initial policy distribution for subsequent RL. This distribution establishes the model's behavioral trajectory and implicitly constrains its exploration space, thereby determining which reasoning paths the model prioritizes during RL.

Motivated by this connection, we revisit the cross-entropy loss for next token prediction. Although traditionally viewed as a supervised metric, cross-entropy can be interpreted as a specific instance of policy gradient optimization within a single-step episode (Wu et al., 2025; Ming et al., 2025). This interpretation suggests that next-token prediction inherently permits an on-policy perspective, even though standard teacher forcing utilizes off-policy samples drawn directly from the training corpus distribution. From an entropy perspective, cross-entropy implicitly assigns maximal reward to the single ground-truth token while uniformly suppressing all negative tokens. Building on this insight, we aim to establish a unified pre-training objective that subsumes cross-entropy as a special case, enabling a systematic study of how reward configurations during pre-training influence subsequent RL dynamics.

In this paper, we propose a generalized objective that integrates on-policy training principles into supervised learning. By formulating next-token prediction as a stochastic decision process, we expose the intrinsic reward mechanism of cross-entropy and introduce a reward-shaping strategy. This approach explicitly regulates the trade-off between diversity and precision during pre-training, rather than deferring this balance to subsequent RL stages. Specifically,

[†] Project Lead.

we introduce a positive reward scaling factor to control the concentration of probability mass on ground-truth tokens, and we differentiate between high-ranking and low-ranking negative tokens to modulate suppression asymmetrically. This strategy allows us to reshape the token output distribution and systematically analyze the relationship between pre-training objectives and RL exploration. Contrary to the conventional intuition that higher distribution entropy facilitates effective exploration, our findings reveal that imposing a precision-oriented prior yields a superior exploration space for RL, ultimately enhancing end-to-end reasoning performance.

Our main contributions are summarized as follows:

- We propose a generalized pre-training objective for next-token prediction that incorporates a reward-shaping strategy, utilizing a positive reward scaling factor and rank-aware negative suppression.
- We investigate how reshaping the token output distribution during pre-training modulates the exploration space for subsequent RL, thereby impacting end-to-end reasoning performance.
- We demonstrate that a precision-oriented pre-training prior provides a more effective initialization for RL than high-entropy distributions, leading to improved reasoning capabilities.

2 Method

2.1 Next Token Prediction

Autoregressive LLMs are typically trained using a next-token prediction objective. This process can be formulated as a sequential decision-making problem where the LLM functions as a stochastic policy π_θ.

Let $X = \{x_1, x_2, \cdots, x_n\}$ denote a sequence of n tokens. At step t, the state s_t is defined by the prefix $X_{<t} = \{x_1, x_2, \cdots, x_{t-1}\}$. The action a_t corresponds to the next token, sampled from the vocabulary V according to the policy $\pi_\theta(\cdot \mid s_t)$. The training objective optimizes the parameters θ to maximize the expected cumulative reward:

$$J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\Big[\sum_{t=1}^{n} r(s_t, a_t)\Big], \tag{1}$$

where $\tau = (s_1, a_1, s_2, a_2, \cdots)$ represents a trajectory sampled from π_θ, and $r(s_t, a_t)$ is the scalar reward received for taking action a_t in state s_t. The policy gradient can be derived as:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\Big[\sum_{t=1}^{n} R(\tau)\nabla_\theta \log \pi_\theta(a_t \mid s_t)\Big], \tag{2}$$

where $R(\tau) = \sum_{t'=1}^{n} r(s_{t'}, a_{t'})$. To reduce variance without introducing bias, the total return $R(\tau)$ is typically replaced by the return-to-go $G_t = \sum_{t'=t}^{n} r(s_{t'}, a_{t'})$, often incorporating a baseline $b(s_t)$ for variance reduction:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\Big[\sum_{t=1}^{n} (G_t - b(s_t))\nabla_\theta \log \pi_\theta(a_t \mid s_t)\Big]. \tag{3}$$

Building upon Equation (3), we treat the generation of a single token as a complete episode (Ming et al., 2025). The objective for a fixed state s_t simplifies to:

$$J_t(\theta \mid s_t) = \mathbb{E}_{a_t\sim\pi_\theta(\cdot|s_t)}[r(s_t, a_t)], \tag{4}$$

yielding the gradient:

$$\nabla_\theta J_t(\theta \mid s_t) = \mathbb{E}_{a_t\sim\pi_\theta(\cdot|s_t)}\big[r(s_t, a_t)\nabla_\theta \log \pi_\theta(a_t \mid s_t)\big]. \tag{5}$$

Crucially, for Equation (5) to remain consistent with the cumulative reward structure of Equation (3), the reward $r(s_t, a_t)$ must depend solely on the immediate state-action pair.

2.2 Revisiting Cross-Entropy

LLM pre-training is generally cast as a supervised learning process designed to maximize the log-likelihood of the ground-truth token x_t given the context $s_t = X_{<t}$:

$$J_{\text{CE}}(\theta) = \log \pi_\theta(x_t \mid s_t). \tag{6}$$

The gradient of this objective explicitly maximizes the probability of the ground-truth token:

$$\nabla_\theta J_{\text{CE}}(\theta) = \nabla_\theta \log \pi_\theta(x_t \mid s_t). \tag{7}$$

We can express this gradient as an expectation over the full policy distribution $\pi_\theta(\cdot \mid s_t)$, encompassing both positive ($a_t = x_t$) and negative ($a_t \neq x_t$) tokens. By invoking the log-derivative identity $\nabla_\theta \log \pi_\theta(x) = \frac{\nabla_\theta \pi_\theta(x)}{\pi_\theta(x)}$

and introducing the indicator function $\mathbb{1}(a_t = x_t)$, we expand the gradient into a summation over the vocabulary V:

$$\begin{aligned}\nabla_\theta J_{\mathrm{CE}}(\theta) &= \frac{1}{\pi_\theta(x_t \mid s_t)} \nabla_\theta \pi_\theta(x_t \mid s_t) \\ &= \frac{1}{\pi_\theta(x_t \mid s_t)} \sum_{a_t \in V} \mathbb{1}(a_t = x_t) \nabla_\theta \pi_\theta(a_t \mid s_t). \end{aligned} \tag{8}$$

We recover the probability density using the substitution $\nabla_\theta \pi_\theta(a_t \mid s_t) = \pi_\theta(a_t \mid s_t) \nabla_\theta \log \pi_\theta(a_t \mid s_t)$, and then form an expectation:

$$\begin{aligned}\nabla_\theta J_{\mathrm{CE}}(\theta) &= \sum_{a_t \in V} \pi_\theta(a_t \mid s_t) \left[\frac{\mathbb{1}(a_t = x_t)}{\pi_\theta(a_t \mid s_t)} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \right] \\ &= \mathbb{E}_{a_t \sim \pi_\theta(\cdot \mid s_t)} \left[r_{\mathrm{CE}}(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t \mid s_t) \right]. \end{aligned} \tag{9}$$

In supervised training, the ground-truth token x_t is deterministically defined by the dataset. Consequently, the indicator $\mathbb{1}(a_t = x_t)$ evaluates the action a_t against a static property of s_t, ensuring that the derived intrinsic reward depends exclusively on information available at step t. Comparing Equation (9) with Equation (5) reveals the intrinsic reward function of cross-entropy:

$$r_{\mathrm{CE}}(s_t, a_t) = \mathrm{sg}\left(\frac{\mathbb{1}(a_t = x_t)}{\pi_\theta(a_t \mid s_t)}\right), \tag{10}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator. Equation (10) demonstrates that when the sampled action matches the ground truth ($a_t = x_t$), the reward is scaled by the inverse probability $\frac{1}{\pi_\theta(x_t \mid s_t)}$. On the contrary, for all negative tokens, the intrinsic reward is exactly 0. Unlike RL scenarios where negative actions are often explicitly penalized, cross-entropy achieves suppression of negative tokens implicitly through the Softmax normalization constraint $\sum\limits_{a_t \in V} \pi_\theta(a_t \mid s_t) = 1$. By increasing the probability of the positive tokens via positive rewards, the probabilities of competing tokens are forced to decrease.

2.3 Diversity or Precision

As derived in Equation (10), the intrinsic reward of the cross-entropy objective implicitly balances diversity and precision. To explicitly regulate the trade-off between these two objectives, we propose a generalized reward function designed to independently control the influence of positive and negative tokens.

First, we introduce a modulating factor to scale the reward associated with the ground-truth token. Let a_t denote the generated token and x_t the ground truth, we define the modified positive reward as:

$$\bar{r}_{\mathrm{pos}}(s_t, a_t) = \mathrm{sg}\left(\left(\frac{1}{\pi_\theta(a_t \mid s_t)}\right)^{(1-\pi_\theta(a_t \mid s_t))^\beta}\right), \tag{11}$$

where $(1 - \pi_\theta(a_t \mid s_t))^\beta$ serves as a positive reward scaling factor. Equation (11) facilitates the control of global entropy. Specifically, when $\beta < 0$, the reward is amplified relative to the baseline ($\beta = 0$). This produces large gradient updates that aggressively concentrate probability mass onto the ground truth, collapsing the distribution and minimizing global entropy. Conversely, $\beta > 0$ attenuates the reward signal. In this regime, the model is less penalized for assigning a lower probability to the ground truth, allowing the policy to maintain a flatter distribution with higher entropy.

Second, while standard cross-entropy assigns zero reward to all negative tokens, we propose shaping the negative distribution to control local entropy. Let $\mathcal{K}_t = \mathrm{TopK}(\pi_\theta(\cdot \mid s_t), k)$ denote the set of the top-k predicted tokens, we define the negative reward as:

$$\bar{r}_{\mathrm{neg}}(s_t, a_t) = \tilde{\lambda} \cdot \mathbb{1}(a_t \in \mathcal{K}_t \wedge a_t \neq x_t) + \hat{\lambda} \cdot \mathbb{1}(a_t \notin \mathcal{K}_t \wedge a_t \neq x_t). \tag{12}$$

As shown in Equation (12), we assign a reward $\tilde{\lambda}$ to high-ranking negative tokens to prevent the model from becoming overly confident in the ground truth alone, thereby reserving probability mass for plausible alternatives. Meanwhile, to suppress low-probability tail tokens, we apply a reward $\hat{\lambda}$ to tokens falling outside $\mathcal{K}_t$, forcing the distribution to concentrate on the head.

Finally, the generalized reward function for the single-step objective is defined as:

$$\bar{r}(s_t, a_t) = \bar{r}_{\mathrm{pos}}(s_t, a_t) \cdot \mathbb{1}(a_t = x_t) + \bar{r}_{\mathrm{neg}}(s_t, a_t) \cdot \mathbb{1}(a_t \neq x_t). \tag{13}$$

Notably, the setting $\beta = 0, \tilde{\lambda} = 0, \hat{\lambda} = 0$ recovers standard cross-entropy.

3 Experiments

3.1 Training Settings

The training pipeline proceeds in three stages: pre-training, mid-training, and RLVR. Adhering to the Qwen3 (Yang et al., 2025), we develop LLMs using both dense and MoE architectures. Specifically, we develop a series of LLMs, which include 1B and 4B dense models, as well as 5B-A0.3B and 10B-A0.5B MoE models. Moreover, we conduct the complete training pipeline on the 4B and 10B-A0.5B models, while the 1B and 5B-A0.3B models undergo the pre-training stage only. More training details are provided in Section A and Section B.

Training Data. For pre-training, we curate a corpus of 500B tokens primarily focused on general knowledge. This is followed by a mid-training stage comprising 100B tokens, which incorporates approximately 5% synthetic data and significantly increases the proportion of reasoning-oriented content. Crucially, we deliberately exclude the synthetic long-reasoning data from all training stages to accurately observe the activation trends of the model's long-CoT reasoning capabilities. The RL stage prioritizes mathematical reasoning tasks, as the emergence of long-reasoning capabilities is typically associated with these domains.

Hyperparameters. Hyperparameters are maintained across the pre-training and mid-training stages. Our goal is to investigate how different reward shaping strategies influence end-to-end performance. Consequently, we perform specific reward configurations for positive tokens ($\beta = -0.25$ and $\beta = 0.5$) and negative tokens ($\hat{\lambda} = -0.1, \tilde{\lambda} = 0, k = 100$ and $\hat{\lambda} = 0, \tilde{\lambda} = 0.1, k = 100$). Employing these distinct hyperparameter configurations allows us to isolate the specific effects of positive and negative reward signals.

3.2 Evaluation Settings

Evaluation of Base Models. Our comprehensive evaluation of base models assesses five core capabilities: general knowledge, logic reasoning, commonsense reasoning, mathematics, and coding. The evaluation is conducted using 19 distinct benchmarks:

- **General Knowledge:** MMLU (Hendrycks et al., 2020)(4-shot, CoT), MMLU-Pro (Wang et al., 2024)(5-shot, CoT), TriviaQA (Joshi et al., 2017)(5-shot), and NaturalQuestions (Kwiatkowski et al., 2019)(5-shot).
- **Commonsense Reasoning:** Hellaswag (Zellers et al., 2019)(0-shot), SIQA (Sap et al., 2019)(0-shot), PIQA (Bisk et al., 2020)(0-shot), WinoGrande (Sakaguchi et al., 2021)(0-shot), OpenBookQA (Mihaylov et al., 2018)(5-shot), and CommonsenseQA (Talmor et al., 2018)(5-shot)
- **Logic Reasoning:** ARC-Easy (Clark et al., 2018)(0-shot), ARC-Challenge (Clark et al., 2018)(0-shot), and BBH (Suzgun et al., 2022)(3-shot, CoT)
- **Mathematics:** GSM8K (Cobbe et al., 2021)(4-shot, CoT), MATH-500 (Lightman et al., 2023)(4-shot, CoT), Minerva (Lewkowycz et al., 2022)(4-shot, CoT), and OlympiadBench (He et al., 2024)(0-shot).
- **Coding:** HumanEval+ (Liu et al., 2023)(0-shot) and MBPP+ (Liu et al., 2023)(3-shot).

Specifically, general knowledge and commonsense reasoning evaluate the model's knowledge-base capabilities, whereas logical reasoning, mathematics, and coding probe its reasoning-base capabilities. Moreover, we employ the Pass@k metric to evaluate the model's upper-bound capability for tasks requiring mathematical reasoning and code generation. Pass@k measures the probability that at least one correct solution is present within k independent attempts. We utilize the unbiased estimator of Pass@k (Chen, 2021), which is defined as:

$$\text{Pass@}k = 1 - \frac{\binom{m-c}{k}}{\binom{m}{k}}, \tag{14}$$

where m represents the total number of sampled responses generated per prompt, and c denotes the count of correct responses among those m samples. We sample $m = 128$ responses with temperature 0.7 and top-p 0.95 and report Pass@64 metric. Notably, we configure the maximum output length to 4K for pre-trained models and 16K for mid-trained models.

Evaluation of RL Models. For RL models evaluation, we employ various mathematics benchmarks, including AMC23 (MAA, b), AIME (MAA, a), MATH-500 (Lightman et al., 2023), Minerva (Lewkowycz et al., 2022), and OlympiadBench (He et al., 2024). We sample 128 responses per problem and report Avg@128, Cons@128, and Pass@64 metrics. Specifically, Avg@128 represents the average accuracy across all 128 samples, while Cons@128 refers to the majority voting accuracy. Similarly, we configure the maximum output length to 16K for RL models.

3.3 Pre-Training

Our analysis of the proposed generalized training objective reveals that it effectively regulates the trade-off between diversity and precision by strategically varying reward configurations. As illustrated in Figure 1 and Figure 2, perplexity (PPL) consistently converges to comparable low values across both dense (1B, 4B) and MoE (5B-A0.3B, 10B-A0.5B) architectures. This demonstrates that, within a specific range, modifying the reward function modulates

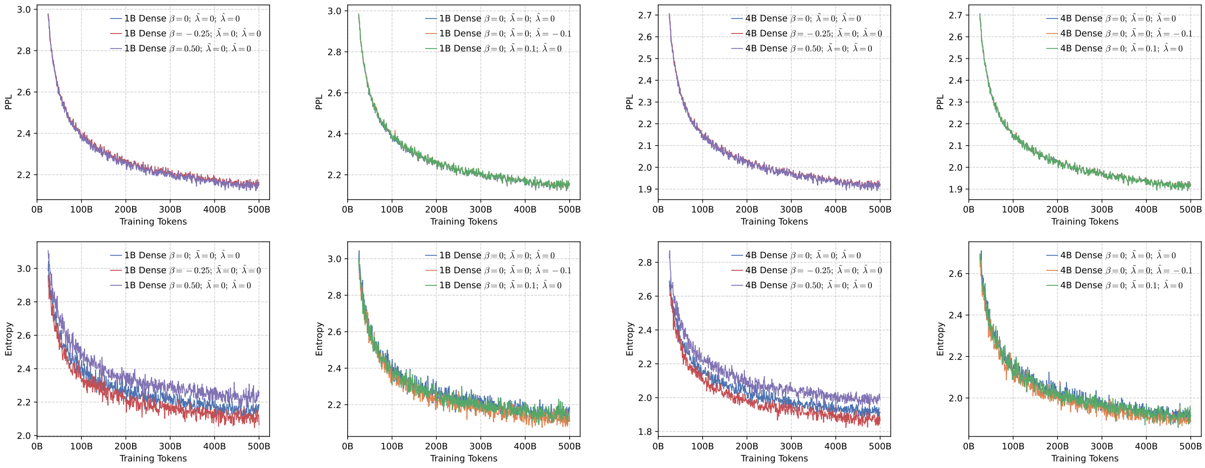

Figure 1: Changes of PPL and entropy during pre-training across 1B and 4B dense models, developed based on different configurations.

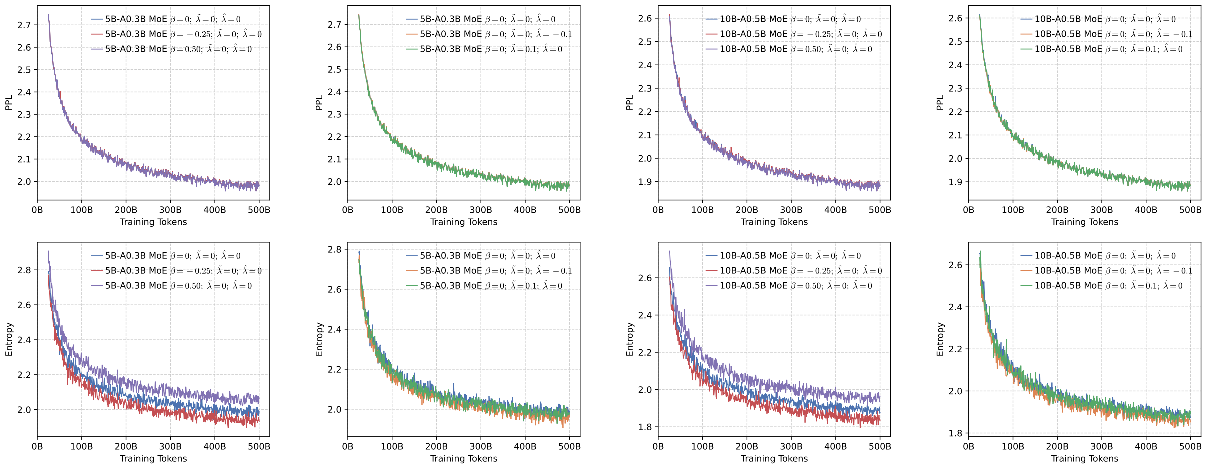

Figure 2: Changes of PPL and entropy during pre-training across 5B-A0.3B and 10B-A0.5B MoE models, developed based on different configurations.

training dynamics without compromising final predictive accuracy. The parameter β serves as a potent global entropy regulator. Specifically, setting $\beta < 0$ significantly reduces entropy, resulting in a more peaked and confident token distribution by amplifying rewards for ground turth tokens. Conversely, $\beta > 0$ maintains higher entropy and a flatter distribution, thereby promoting diversity in the generated output. Meanwhile, the parameters $\hat{\lambda}$ and $\tilde{\lambda}$ facilitate local entropy fine-tuning. These parameters shape the token distribution by either rewarding ($\hat{\lambda} = 0, \tilde{\lambda} = 0.1, k = 100$) or penalizing ($\hat{\lambda} = -0.1, \tilde{\lambda} = 0, k = 100$) negative tokens, enabling granular control over the training process.

Furthermore, we analyze the evolution of model performance during pre-training to investigate the dynamics and specific impact of the proposed reward function. As depicted in Figure 3, larger models consistently achieve substantially higher final performance than smaller models after processing an equivalent number of training tokens. This confirms that explicitly regulating the diversity-precision trade-off is an orthogonal mechanism that does not interfere with the fundamental scaling properties of language models. Crucially, configurations that prioritize lowering global entropy ($\beta < 0$) or maintaining high local entropy ($\hat{\lambda} = -0.1, \tilde{\lambda} = 0, k = 100$) demonstrate superior performance and scaling behavior. Although these settings may not yield optimal initial performance in smaller models, they exhibit enhanced growth potential as model size increases. This suggests that with greater model capacity, strategies that promote precision, either globally via generously rewarding positive tokens or locally by aggressively penalizing tail negative tokens, lead to better performance growth compared to the baseline.

3.4 Mid-Training

Subsequently, we evaluate the evolution of model performance during the mid-training stage, spanning from 0B to 100B tokens. As depicted in Figure 4, the choice of β significantly influences training dynamics. We observe a consistent trend where a negative value, specifically $\beta = -0.25$, yields the best results. This configuration consistently outperforms the baseline ($\beta = 0$) across both dense and MoE models in knowledge and reasoning

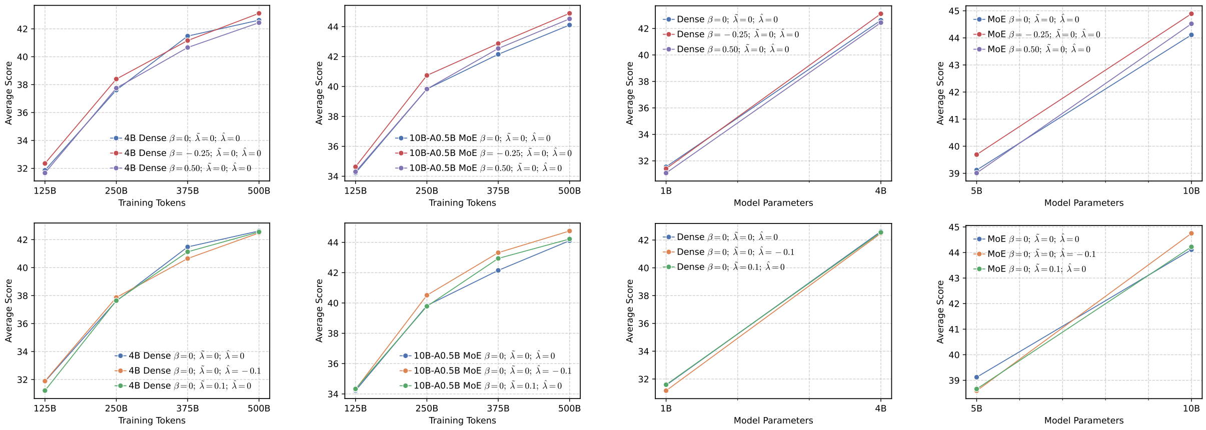

Figure 3: Changes of performance during pre-training across models with various model parameters, developed based on dense and MoE architectures under different configurations.

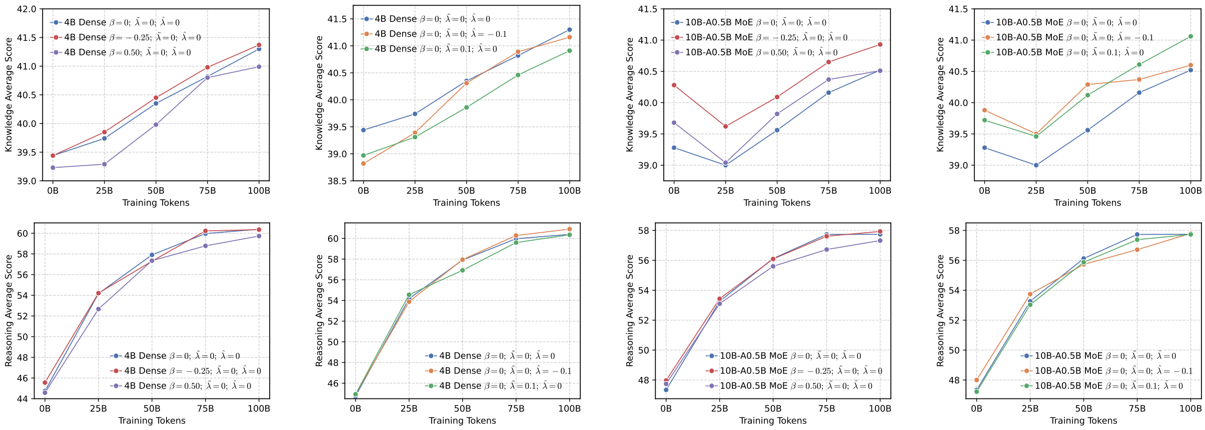

Figure 4: Changes of performance during mid-training across 4B dense and 10B-A0.5B MoE models, developed based on different configurations.

tasks. Conversely, a positive setting ($\beta = 0.50$) does not demonstrate consistent superior performance comparing to the baseline. Similar to the observations with β, a slight negative adjustment appears beneficial. The configuration $\hat{\lambda} = -0.1, \tilde{\lambda} = 0, k = 100$ generally matches or slightly surpasses the performance of the standard CE baseline. However, when shifting to $\hat{\lambda} = 0.1, \tilde{\lambda} = 0, k = 100$, performance exhibited uncertainty in knowledge-intensive scenarios. In reasoning tasks, the performance remained comparable to the standard CE baseline.

3.5 Reinforcement Learning

Finally, we investigate the performance dynamics during the RL training stage across various actor models, as illustrated in Figure 5 and Figure 6. Pre-trained models derived from different reward configurations exhibit distinct output distributions, leading to significant variations in subsequent RL and end-to-end reasoning performance. Regarding the global entropy regulator β, we observe a consistent and robust trend across both the 4B dense and 10B-A0.5B MoE models. Specifically, the global low entropy setting ($\beta = -0.25$) yields superior performance trajectories. This configuration consistently outperforms the global high entropy setting across all evaluated metrics, including Avg@128, Cons@128, and Pass@64. Furthermore, the configuration $\hat{\lambda} = -0.1, \tilde{\lambda} = 0, k = 100$ demonstrates a significant advantage, consistently achieving the highest performance and notably surpassing the baseline on the 10B-A0.5B MoE model. For the 4B dense model, maintaining local high entropy exhibits a superior scaling trend compared to the baseline. In conclusion, strategies that promote precision, either globally via generously rewarding positive tokens or locally by aggressively penalizing tail negative tokens, enables the model to converge to higher-quality solutions, potentially providing a better exploration space for RL.

To better understand the performance divergence observed during RL, we analyze the evolution of policy entropy and response length throughout the training process, as illustrated in Figure 7. Contrary to the expectation that higher entropy maintains diversity, setting a higher β leads to rapid entropy collapse during the early stages of training. Coinciding with this collapse, the response length decreases drastically, indicating a suppression of the reasoning capability. In contrast, local high-entropy configurations exhibit greater stability. These settings effectively prevent entropy collapse, maintaining a robust policy distribution from the onset. They demonstrate a smooth and continuous

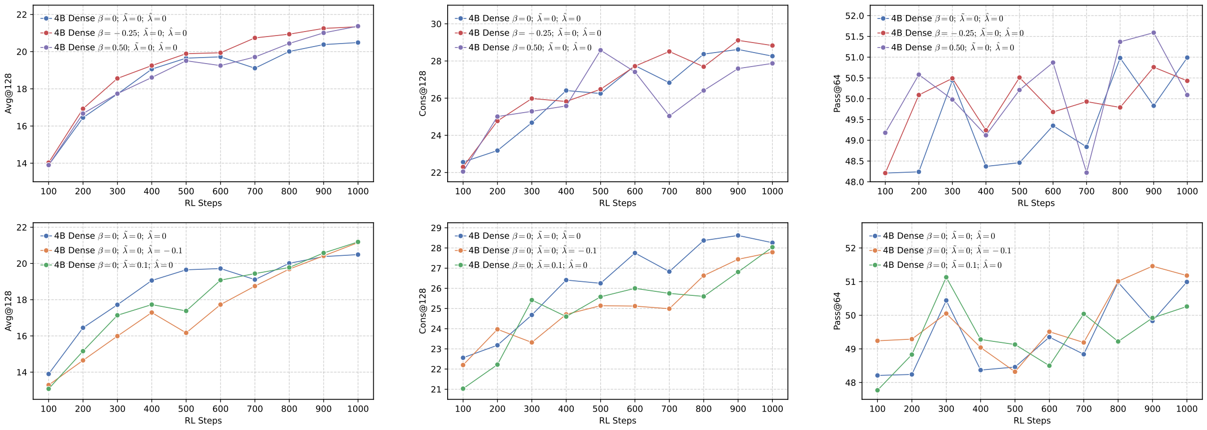

Figure 5: Changes of performance during RL training across various actor models, developed based on a 4B dense architecture under different configurations.

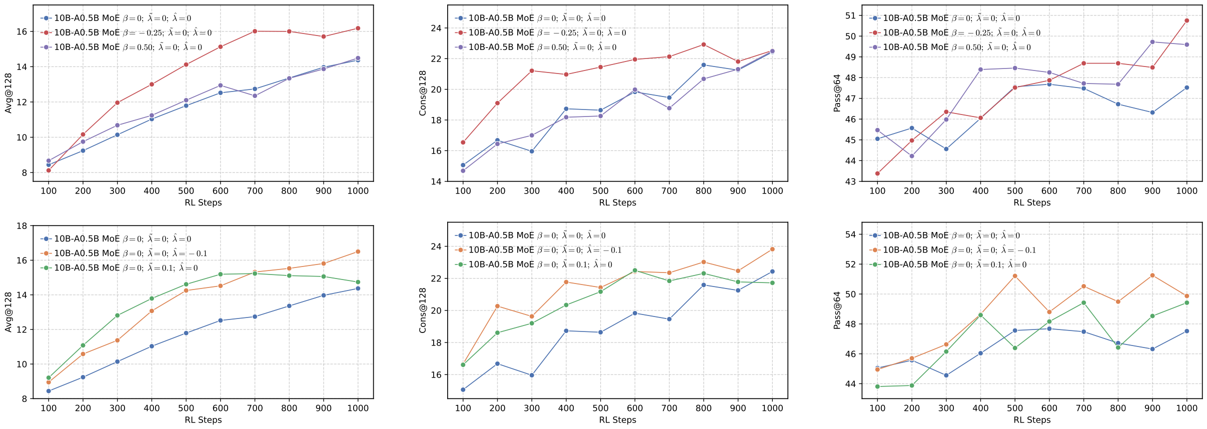

Figure 6: Changes of performance during RL training across various actor models, developed based on a 10B-A0.5B MoE architecture under different configurations.

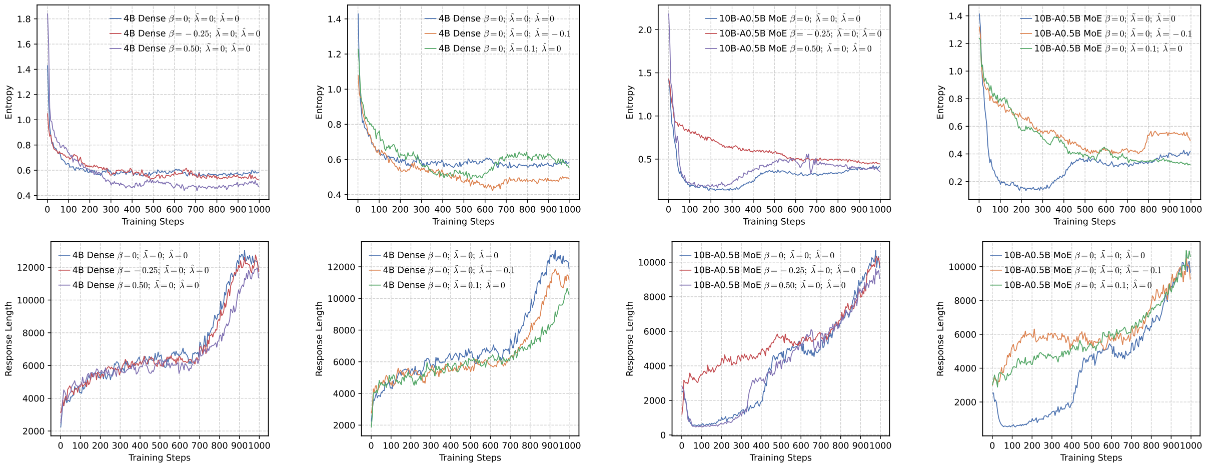

Figure 7: Changes of entropy and response length during RL training across various actor models, developed based on 4B dense and 10B-A0.5B MoE architectures under different configurations.

activation of long reasoning capabilities, allowing for a steady increase in generation length and reasoning depth without the recovery lag observed in global high entropy settings.

3.6 Pass@k Analysis of Base Models

Moreover, we analyze the Pass@k curves as k increases to estimate the upper bound of the capability of base models. This metric relies on a delicate equilibrium between solution precision and diversity. As shown in Figure 8, maximizing global diversity (high entropy) does not inherently yield higher Pass@k curves. Instead, superior

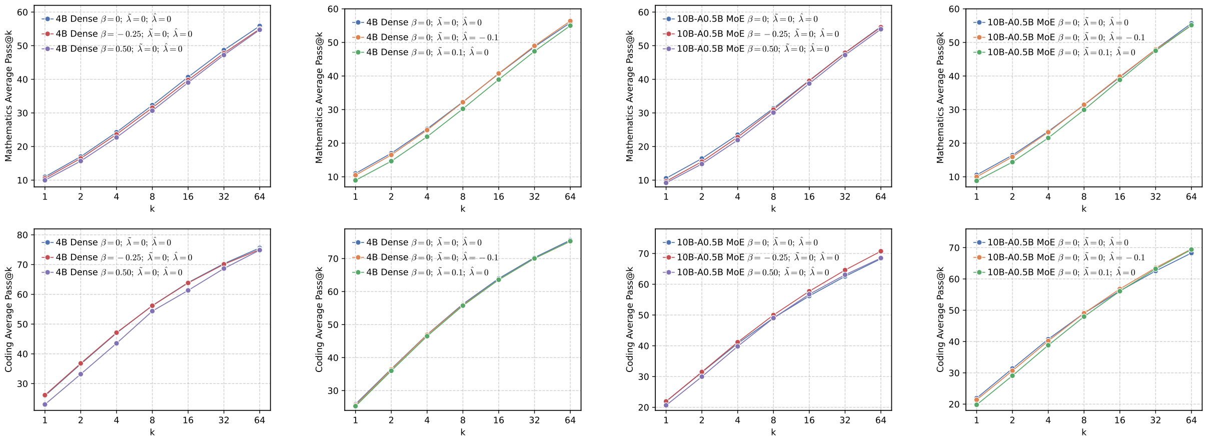

Figure 8: Pass@k curve of base models on mathematics reasoning and code generation tasks, developed based on 4B dense and 10B-A0.5B MoE models under different configurations.

Pass@k scores in mathematics and coding tasks are achieved by prioritizing precision. Crucially, we observe that this low-entropy setting does not lead to a collapse in output diversity. Rather, it maintains sufficient variation to cover the solution space. Furthermore, the data indicate that promoting local diversity also yields better results. This suggests that while models benefit from high precision, they simultaneously benefit from targeted local exploration.

4 Related Works

4.1 Weighted Cross-Entropy Loss

The standard cross-entropy objective can be generalized within a policy-gradient framework, where it is equivalent to optimizing a sparse reward defined as $r_{\mathrm{CE}}(s_t, a_t) = \mathbb{1}(a_t = x_t)\pi_\theta(a_t \mid s_t)^{-1}$. Existing modifications to this objective include smooth loss (label smoothing), which encourages diversity by allocating a uniform probability mass to all positive tokens, and focal loss (Lin et al., 2018), which down-weights easy examples via $w_t = (1 - \pi_\theta(x_t \mid s_t))^\gamma$. Our proposed generalized training objective can also formulate these established variations. In this paper, we specifically explore two different reward configurations within this framework. Firstly, we introduce a modified positive reward, which is equivalent to applying a state-dependent weight $w_t = \pi_\theta(x_t \mid s_t)^{1-(1-\pi_\theta(x_t|s_t))^\beta}$ to the standard cross-entropy. In addition, we incorporate TopK-based negative shaping, which explicitly controls local entropy by assigning non-zero rewards to selected actions with $a_t \neq x_t$.

4.2 Next Token Reasoning

Treating each token emission as a distinct episode ensures that the reward depends only on the immediate state-action pair $(s_t,\ a_t)$, thereby preserving unbiased credit assignment. The framework is naturally compatible with architectures that perform iterative internal computation prior to token emission, including latent-reasoning models (Zelikman et al., 2024) and loop transformers (Dehghani et al., 2019; Zhu et al., 2025a). Although each episode terminates at token emission, the state s_t may encode the outcome of internal refinement cycles. Our reward design can serve as an uncertainty-aware learning signal that can be combined with adaptive computation policies to allocate additional internal processing steps in uncertain contexts. By explicitly shaping positive and negative token-level rewards within a single-step policy-gradient framework, we provide a general and controllable mechanism that natively supports reasoning-oriented architectures through principled reward design.

5 Conclusion

This study establishes a theoretical bridge between next-token prediction and RL by interpreting cross-entropy loss as a specific instance of policy gradient optimization. To exploit this connection, we introduce a generalized pre-training objective that utilizes a reward-shaping strategy with positive scaling and rank-aware negative rewards. Our experiments across multiple architectures and scales reveal that regulating the diversity-precision trade-off during pre-training modulates token entropy. Our findings indicate that precision-focused strategies (e.g., global entropy reduction or tail-token suppression) yield superior scaling for the subsequent RL stage. These insights provide a novel perspective on optimizing pre-training for long CoT reasoning, suggesting new directions for sophisticated reward shaping in LLM development.

References

Anthropic. Claude Sonnet. https://www.anthropic.com/claude/sonnet, 2025.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. PIQA: Reasoning about Physical Commonsense in Natural Language. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pp. 7432–7439, 2020.

Mark Chen. Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374*, 2021.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think You Have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The Entropy Mechanism of Reinforcement Learning for Reasoning Language Models. *arXiv preprint arXiv:2505.22617*, 2025.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal Transformers. *arXiv preprint arXiv:1807.03819*, 2019.

Qingxiu Dong, Li Dong, Yao Tang, Tianzhu Ye, Yutao Sun, Zhifang Sui, and Furu Wei. Reinforcement Pre-Training. *arXiv preprint arXiv:2506.08007*, 2025.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D. Goodman. Cognitive Behaviors that Enable Self-Improving Reasoners, or, Four Habits of Highly Effective STaRs. *arXiv preprint arXiv:2503.01307*, 2025.

Google DeepMind. Gemini3 Pro. https://deepmind.google/technologies/gemini/pro/, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1: Incentivizes Reasoning in LLMs through Reinforcement Learning. *Nature*, 645(8081):633–638, 2025.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems. *arXiv preprint arXiv:2402.14008*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

Kimi, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. Kimi K2: Open Agentic Intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving Quantitative Reasoning Problems with Language Models. In *Annual Conference on Neural Information Processing Systems (NIPS)*, volume 35, pp. 3843–3857, 2022.

Siheng Li, Kejiao Li, Zenan Xu, Guanhua Huang, Evander Yang, Kun Li, Haoyuan Wu, Jiajia Wu, Zihao Zheng, Chenchen Zhang, et al. Reinforcement Learning on Pre-Training Data. *arXiv preprint arXiv:2509.19249*, 2025.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. In *International Conference on Learning Representations (ICLR)*, 2023.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. *arXiv preprint arXiv:1708.02002*, 2018.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. DeepSeek-V3 Technical Report. *arXiv preprint arXiv:2412.19437*, 2024.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is Your Code Generated by ChatGPT Really Correct? Rigorous Evaluation of Large Language Models for Code Generation. In *Annual Conference on Neural Information Processing Systems (NIPS)*, volume 36, pp. 21558–21572, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101*, 2017.

MAA. American Invitational Mathematics Examination (AIME), Mathematics Competition Series, a. URL https://maa.org/math-competitions/aime.

MAA. American Mathematics Competitions (AMC 10/12), Mathematics Competition Series, b. URL https://maa.org/math-competitions/amc.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can A Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. *arXiv preprint arXiv:1809.02789*, 2018.

Rui Ming, Haoyuan Wu, Shoubo Hu, Zhuolun He, and Bei Yu. One-Token Rollout: Guiding Supervised Fine-Tuning of LLMs with Policy Gradient. *arXiv preprint arXiv:2509.26313*, 2025.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An Adversarial Winograd Schema Challenge at Scale. *Communications of the ACM*, 64(9):99–106, 2021.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. SocialiQA: Commonsense Reasoning about Social Interactions. *arXiv preprint arXiv:1904.09728*, 2019.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced Transformer with Rotary Position Embedding. *Neurocomputing*, 2024.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging Big-Bench Tasks and Whether Chain-of-Thought Can Solve Them. *arXiv preprint arXiv:2210.09261*, 2022.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. *arXiv preprint arXiv:1811.00937*, 2018.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 Rule: High-Entropy Minority Tokens Drive Effective Reinforcement Learning for LLM Reasoning. *arXiv preprint arXiv:2506.01939*, 2025.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. In *Annual Conference on Neural Information Processing Systems (NIPS)*, 2024.

Yongliang Wu, Yizhou Zhou, Zhou Ziheng, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. On the Generalization of SFT: A Reinforcement Learning Perspective with Reward Rectification. *arXiv preprint arXiv:2508.05629*, 2025.

Xingrun Xing, Zhiyuan Fan, Jie Lou, Guoqi Li, Jiajun Zhang, and Debing Zhang. PretrainZero: Reinforcement Active Pretraining. *arXiv preprint arXiv:2512.03442*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. *arXiv preprint arXiv:2503.14476*, 2025.

Eric Zelikman, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D Goodman. Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking. *arXiv preprint arXiv:2403.09629*, 2024.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can A Machine Really Finish Your Sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Rui-Jie Zhu, Zixuan Wang, Kai Hua, Tianyu Zhang, Ziniu Li, Haoran Que, Boyi Wei, Zixin Wen, Fan Yin, He Xing, Lu Li, Jiajun Shi, Kaijing Ma, Shanda Li, Taylor Kergan, Andrew Smith, Xingwei Qu, Mude Hui, Bohong Wu, Qiyang Min, Hongzhi Huang, Xun Zhou, Wei Ye, Jiaheng Liu, Jian Yang, Yunfeng Shi, Chenghua Lin, Enduo Zhao, Tianle Cai, Ge Zhang, Wenhao Huang, Yoshua Bengio, and Jason Eshraghian. Scaling Latent Reasoning via Looped Language Models. *arXiv preprint arXiv:2510.25741*, 2025a.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The Surprising Effectiveness of Negative Reinforcement in LLM Reasoning. *arXiv preprint arXiv:2506.01347*, 2025b.

A Experiment Details for Pre-Training and Mid-Training

A.1 Implementation Details

For both the pre-training and mid-training phases, we employ the AdamW (Loshchilov & Hutter, 2017) optimizer, implementing a weight decay of 0.1 and applying gradient clipping at 1.0. Throughout these stages, we utilize a warmup-stable-decay learning rate schedule with a global batch size of 16M. During the stable pre-training stage, which encompasses 500B tokens, the learning rate warms up over 2000 steps before stabilizing at 3×10^{-4}. Subsequently, we perform mid-training on an additional 100B tokens, gradually decaying the learning rate from 3×10^{-4} to 3×10^{-5}. We set the maximum sequence length to 4096 during pre-training and extend it to 16384 for the mid-training stage. To support long-context modeling during mid-training, we increase the base frequency of RoPE (Su et al., 2024) from $1e^4$ to $1e^6$.

A.2 Model Architecture

Building upon the Qwen3 (Yang et al., 2025) architectures, we perform our experiments utilizing both dense and MoE architectures. Notably, we adopt an auxiliary loss free approach (Liu et al., 2024) for the training of the MoE models. Detailed architecture settings are provided in Table 1, where E denotes the total number of experts and E_a denotes the number of active experts.

Table 1: **Detailed architectures settings of dense and MoE models.**

Model	n_{layer}	d_{model}	d_{ffn}	d_{expert}	n_{head}	n_{kvhead}	E	E_a
1B Dense	28	1536	4608	-	16	4	-	-
4B Dense	36	2560	9728	-	32	8	-	-
5B-A0.3B MoE	12	1024	-	320	32	4	384	12
10B-A0.5B MoE	16	1536	-	320	32	4	384	12

A.3 Experiment Results

We report comprehensive evaluation results to demonstrate performance progression throughout the training process. Tables 2 to 9 present the pre-training results across various models and different training tokens. Similarly, Tables 10 to 13 summarize the performance metrics for the mid-training stage.

B Experiment Details for RL

B.1 Implementation Details

For RLVR on mathematical reasoning tasks, we employ the on-policy GRPO algorithm (Shao et al., 2024) without KL regularization. Following Yu et al. (2025), we incorporate clip-higher and dynamic sampling strategies to stabilize training. The process is conducted in two stages: an initial 700 steps with a sequence length of 8K, followed by continued training at a sequence length of 16K. We maintain a batch size of 128 and a constant learning rate of 1×10^{-6} for two stages. During training, we sample 16 outputs per prompt at a temperature of 1.0.

B.2 Experiment Results

We provide detailed evaluation results to illustrate performance trajectories during the RL process. Tables 14 to 23 display the RL results across different models and training steps.

Table 2: **Pre-Training performance comparison across different β based on 1B dense models. The highest scores at the final checkpoint across the different configurations are shown in bold.**

	Hyperparameters	$\beta = -0.25;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$				$\beta = 0;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$				$\beta = 0.50;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$			
	# Pre-Trained Tokens	125B	250B	375B	500B	125B	250B	375B	500B	125B	250B	375B	500B
General Knowledge	MMLU (Acc.)	20.47	23.65	24.06	25.12	20.85	23.39	24.04	**26.62**	21.71	24.26	24.37	25.51
	MMLU-Pro (Acc.)	8.01	9.25	9.54	**10.06**	7.63	8.23	7.98	9.02	7.85	8.55	9.44	9.59
	NaturalQuestions (EM)	3.02	4.34	4.88	5.35	2.74	4.07	4.52	**5.79**	2.69	3.77	4.18	5.21
	TriviaQA (EM)	8.45	12.32	14.75	16.03	8.65	12.63	13.53	16.25	8.35	12.39	14.24	**16.49**
	Average	9.99	12.39	13.31	14.14	9.97	12.08	12.52	**14.42**	10.15	12.24	13.06	14.20
Commonsense Reasoning	Hellaswag (Acc.)	38.24	44.17	46.17	47.49	38.96	44.42	46.26	**48.33**	38.88	43.95	46.83	48.06
	SIQA (Acc.)	38.43	42.22	40.53	39.15	40.02	40.33	41.50	**42.32**	39.36	40.84	42.02	**42.32**
	PIQA (Acc.)	67.36	69.53	69.70	**71.27**	67.90	69.75	71.00	71.11	67.57	69.59	70.62	70.02
	WinoGrande (Acc.)	51.70	49.96	52.09	51.62	52.49	52.17	52.80	**53.83**	53.43	54.54	54.70	53.28
	OpenBookQA (Acc.)	31.40	32.40	31.80	32.20	29.80	32.20	33.60	32.80	30.00	31.60	31.80	**33.40**
	CommonsenseQA (Acc.)	19.66	19.00	21.13	**20.80**	19.57	21.79	19.82	20.07	20.15	19.41	20.72	20.56
	Average	41.13	42.88	43.57	43.76	41.46	43.44	44.16	**44.74**	41.57	43.32	44.45	44.61
Knowledge Average		25.56	27.64	28.44	28.95	25.71	27.76	28.34	**29.58**	25.86	27.78	28.75	29.40
Logic Reasoning	ARC-e (Acc.)	52.74	55.30	59.43	**59.26**	51.64	55.68	56.70	58.80	49.24	55.64	55.51	58.88
	ARC-c (Acc.)	25.34	27.47	30.03	**30.97**	27.30	29.44	29.52	29.01	25.51	27.05	27.30	27.90
	BBH (Acc.)	23.47	23.41	26.46	26.68	22.13	22.56	25.68	**27.34**	25.13	22.49	24.99	26.37
	Average	33.85	35.39	38.64	**38.97**	33.69	35.89	37.30	38.38	33.29	35.06	35.93	37.72
Mathematics	GSM8K (Pass@64)	40.06	43.01	46.15	49.52	41.09	48.78	48.77	48.76	43.52	46.65	46.58	**49.98**
	MATH-500 (Pass@64)	34.16	37.09	39.79	38.15	33.41	35.43	37.10	38.49	30.41	39.64	40.80	**38.97**
	Minerva (Pass@64)	14.62	16.19	17.07	15.76	16.44	16.71	14.80	**16.43**	14.94	15.32	16.54	15.99
	OlympiadBench (Pass@64)	20.39	22.72	23.50	22.35	21.06	21.01	21.85	21.83	20.83	22.91	23.00	**22.91**
	Average	27.31	29.75	31.63	31.45	28.00	30.48	30.63	31.38	27.43	31.13	31.73	**31.96**
Coding	HumanEval+ (Pass@64)	8.06	13.03	15.46	15.81	8.71	12.15	12.72	**15.95**	7.68	11.92	13.67	14.55
	MBPP+ (Pass@64)	21.39	33.69	40.36	**41.70**	20.63	34.37	37.53	41.67	16.58	31.37	38.18	39.26
	Average	14.73	23.36	27.91	28.76	14.67	23.26	25.13	**28.81**	12.13	21.65	25.93	26.91
Reasoning Average		25.29	29.50	32.73	**33.06**	25.45	29.88	31.02	32.86	24.28	29.28	31.20	32.19
Average		25.40	28.76	31.01	31.41	25.56	29.03	29.95	**31.55**	24.91	28.68	30.22	31.08

Table 3: **Pre-Training performance comparison across different β based on 4B dense models. The highest scores at the final checkpoint across the different configurations are shown in bold.**

	Hyperparameters	$\beta = -0.25;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$				$\beta = 0;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$				$\beta = 0.50;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$			
	# Pre-Trained Tokens	125B	250B	375B	500B	125B	250B	375B	500B	125B	250B	375B	500B
General Knowledge	MMLU (Acc.)	24.98	30.94	34.38	36.38	25.76	29.91	33.61	36.17	24.45	30.14	32.84	**36.55**
	MMLU-Pro (Acc.)	8.81	11.11	11.99	**13.47**	9.25	10.90	11.71	12.23	8.48	9.93	10.96	12.63
	NaturalQuestions (EM)	6.87	9.11	10.89	12.58	6.65	9.75	10.89	12.22	6.32	8.92	10.28	**11.69**
	TriviaQA (EM)	17.26	25.14	30.39	33.39	17.91	25.94	27.80	**33.83**	17.19	24.26	30.10	32.90
	Average	14.48	19.08	21.91	**23.96**	14.89	19.13	21.00	23.61	14.11	18.31	21.05	23.44
Commonsense Reasoning	Hellaswag (Acc.)	50.36	57.81	60.80	**63.01**	50.22	57.00	60.50	62.91	39.15	57.49	60.98	62.87
	SIQA (Acc.)	42.99	44.27	44.32	**45.44**	42.37	43.35	43.65	44.73	41.40	41.71	43.24	44.78
	PIQA (Acc.)	71.98	74.65	75.52	75.57	72.42	74.70	75.35	**76.28**	71.76	74.43	74.32	75.46
	WinoGrande (Acc.)	52.72	56.12	57.22	**59.04**	53.67	55.80	56.67	58.72	53.67	55.80	56.67	58.72
	OpenBookQA (Acc.)	33.00	36.00	36.00	36.80	33.40	36.40	37.00	36.00	34.00	36.00	37.80	**39.40**
	CommonsenseQA (Acc.)	21.13	29.48	37.43	49.63	20.39	28.26	47.91	**52.91**	18.43	29.32	37.10	48.98
	Average	45.36	49.72	51.88	54.92	45.41	49.25	53.51	**55.26**	42.85	48.88	51.91	55.03
Knowledge Average		29.92	34.40	36.90	**39.44**	30.15	34.19	37.26	**39.44**	28.48	33.59	36.48	39.23
Logic Reasoning	ARC-e (Acc.)	61.15	66.84	69.44	**70.29**	60.27	63.05	67.97	67.55	58.96	64.98	65.82	66.75
	ARC-c (Acc.)	31.83	35.49	36.77	**37.80**	30.55	33.96	36.77	37.54	31.91	33.61	37.12	37.12
	BBH (Acc.)	26.14	26.32	30.36	**31.65**	26.54	27.42	29.53	28.29	25.02	26.95	28.17	29.78
	Average	39.71	42.88	45.52	**46.58**	39.12	41.48	44.76	44.46	38.63	41.85	43.70	44.55
Mathematics	GSM8K (Pass@64)	49.66	62.66	67.23	71.19	48.43	60.24	68.75	71.26	50.39	61.71	69.88	**71.98**
	MATH-500 (Pass@64)	39.07	47.73	48.62	51.14	38.19	48.15	48.88	51.54	40.86	46.72	48.23	**51.67**
	Minerva (Pass@64)	15.48	20.73	19.37	20.07	15.09	19.22	19.68	**20.63**	16.33	18.45	17.64	19.85
	OlympiadBench (Pass@64)	22.24	24.87	24.08	24.18	21.79	25.53	23.70	24.84	22.23	23.81	25.33	**24.87**
	Average	31.61	39.00	39.83	41.65	30.88	38.29	40.25	42.07	32.45	37.67	40.27	**42.09**
Coding	HumanEval+ (Pass@64)	17.11	22.94	27.79	**31.29**	17.32	23.22	30.08	29.13	16.73	24.36	27.72	28.52
	MBPP+ (Pass@64)	44.07	59.66	65.54	65.63	40.66	56.64	65.69	**66.29**	43.90	59.74	65.07	65.65
	Average	30.59	41.30	46.67	**48.46**	28.99	39.93	47.89	47.71	30.32	42.05	46.40	47.09
Reasoning Average		33.97	41.06	44.00	**45.56**	33.00	39.90	44.30	44.75	33.80	40.52	43.46	44.58
Average		32.35	38.40	41.16	**43.11**	31.86	37.61	41.48	42.62	31.67	37.75	40.66	42.44

Table 4: **Pre-Training performance comparison across different β based on 5B-A0.3B MoE models. The highest scores at the final checkpoint across the different configurations are shown in bold.**

	Hyperparameters	$\beta = -0.25;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$				$\beta = 0;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$				$\beta = 0.50;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$			
	# Pre-Trained Tokens	125B	250B	375B	500B	125B	250B	375B	500B	125B	250B	375B	500B
General Knowledge	MMLU (Acc.)	22.70	27.21	29.36	29.77	23.10	25.12	29.17	**31.19**	24.13	26.99	28.40	30.34
	MMLU-Pro (Acc.)	8.80	8.56	9.42	11.16	9.47	8.89	10.06	**11.87**	8.62	9.83	9.15	10.36
	NaturalQuestions (EM)	5.21	7.45	8.45	10.06	5.15	7.51	8.73	**10.42**	5.76	7.48	8.78	10.17
	TriviaQA (EM)	15.05	22.23	25.51	28.08	14.93	22.03	25.27	**28.25**	13.75	21.27	25.96	28.00
	Average	12.94	16.36	18.19	19.77	13.16	15.89	18.31	**20.43**	13.07	16.39	18.07	19.72
Commonsense Reasoning	Hellaswag (Acc.)	47.89	54.08	56.70	57.78	48.54	54.63	56.88	57.41	48.58	54.31	56.81	**58.13**
	SIQA (Acc.)	40.53	41.25	42.68	43.14	41.25	42.27	43.76	42.68	40.28	42.43	42.01	**43.19**
	PIQA (Acc.)	71.49	72.63	73.67	**75.24**	71.27	73.18	74.43	74.76	71.60	73.45	74.59	74.59
	WinoGrande (Acc.)	52.88	53.35	57.54	**57.38**	50.43	53.20	56.51	56.20	52.01	54.22	55.09	56.43
	OpenBookQA (Acc.)	30.80	35.60	34.00	**36.60**	31.20	32.80	34.40	33.60	33.80	34.00	35.80	35.60
	CommonsenseQA (Acc.)	20.64	27.27	33.25	**38.57**	18.67	22.93	28.50	35.38	19.41	22.69	25.88	31.37
	Average	44.04	47.36	49.64	**51.45**	43.56	46.50	49.08	50.01	44.28	46.85	48.36	49.89
Knowledge Average		28.49	31.86	33.91	**35.61**	28.36	31.19	33.69	35.22	28.67	31.62	33.22	34.80
Logic Reasoning	ARC-e (Acc.)	57.79	60.98	62.12	62.08	58.54	62.92	64.48	63.34	58.12	64.02	63.55	**63.38**
	ARC-c (Acc.)	27.47	32.25	32.94	34.39	30.80	34.81	35.67	**35.15**	28.67	33.36	34.13	34.56
	BBH (Acc.)	23.30	26.80	27.12	**27.85**	23.97	25.79	27.71	27.03	25.94	26.69	27.00	27.49
	Average	36.19	40.01	40.73	41.44	37.77	41.17	42.62	**41.84**	37.58	41.36	41.56	41.81
Mathematics	GSM8K (Pass@64)	52.99	59.43	64.26	**66.78**	51.48	59.15	61.81	65.65	51.36	56.21	62.12	65.04
	MATH-500 (Pass@64)	39.15	43.06	48.46	**52.00**	37.57	43.69	46.91	51.19	38.99	42.89	47.80	48.26
	Minerva (Pass@64)	17.91	18.98	18.79	**21.79**	15.47	16.93	18.00	17.92	14.81	18.87	17.61	19.36
	OlympiadBench (Pass@64)	22.79	23.65	24.29	25.62	23.50	24.14	23.40	24.42	23.73	23.10	25.18	**26.00**
	Average	33.21	36.28	38.95	**41.55**	32.01	35.98	37.53	39.80	32.22	35.27	38.18	39.67
Coding	HumanEval+ (Pass@64)	18.11	23.31	26.20	28.44	17.52	22.79	26.15	**29.32**	17.14	22.20	27.53	27.95
	MBPP+ (Pass@64)	37.25	55.06	58.57	**60.06**	39.81	54.29	55.93	57.77	38.75	52.50	56.45	59.97
	Average	27.68	39.19	42.39	**44.25**	28.67	38.54	41.04	43.55	27.95	37.35	41.99	43.96
Reasoning Average		32.36	38.49	40.69	**42.41**	32.81	38.56	40.40	41.73	32.58	37.99	40.58	41.81
Average		30.81	35.84	37.98	**39.69**	31.03	35.62	37.72	39.12	31.02	35.44	37.63	39.01

Table 5: **Pre-Training performance comparison across different β based on 10B-A0.5B MoE models. The highest scores at the final checkpoint across the different configurations are shown in bold.**

	Hyperparameters	$\beta = -0.25;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$				$\beta = 0;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$				$\beta = 0.50;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$			
	# Pre-Trained Tokens	125B	250B	375B	500B	125B	250B	375B	500B	125B	250B	375B	500B
General Knowledge	MMLU (Acc.)	26.13	32.74	35.09	**36.60**	25.35	29.14	33.50	35.12	26.05	30.71	34.03	35.48
	MMLU-Pro (Acc.)	9.44	10.21	11.38	**12.26**	8.24	10.01	11.64	12.15	8.81	11.32	10.97	11.90
	NaturalQuestions (EM)	6.79	10.36	12.47	**13.99**	7.06	10.47	11.41	11.96	7.84	10.58	11.36	13.77
	TriviaQA (EM)	20.69	29.78	35.25	37.95	19.94	29.86	33.89	37.31	20.73	29.64	35.31	**39.13**
	Average	15.76	20.77	23.55	**25.20**	15.15	19.87	22.61	24.14	15.86	20.56	22.92	25.07
Commonsense Reasoning	Hellaswag (Acc.)	52.86	59.47	61.65	63.26	53.00	59.27	62.00	**63.71**	53.09	59.17	61.96	63.48
	SIQA (Acc.)	41.61	43.45	45.14	**45.60**	42.02	42.73	43.45	45.09	41.91	42.84	44.63	43.76
	PIQA (Acc.)	73.45	74.32	74.76	75.68	72.74	74.59	75.52	76.71	74.10	74.92	76.55	**76.88**
	WinoGrande (Acc.)	54.06	56.59	56.27	58.64	52.88	56.12	59.04	58.88	53.83	56.04	58.80	**59.04**
	OpenBookQA (Acc.)	34.40	35.00	35.40	37.40	32.80	35.60	35.60	38.20	33.20	37.60	37.20	**39.40**
	CommonsenseQA (Acc.)	20.07	34.64	44.80	**51.60**	22.52	33.74	35.54	43.90	21.54	32.02	37.51	43.16
	Average	46.08	50.58	53.00	**55.36**	45.99	50.34	51.86	54.42	46.28	50.43	52.78	54.29
Knowledge Average		30.92	35.68	38.28	**40.28**	30.57	35.11	37.23	39.28	31.07	35.50	37.85	39.68
Logic Reasoning	ARC-e (Acc.)	60.94	65.91	66.84	**70.29**	63.01	67.09	67.63	67.00	62.37	66.37	68.77	69.32
	ARC-c (Acc.)	31.91	37.20	35.67	37.46	32.68	35.24	36.52	37.29	32.51	36.09	37.20	**38.82**
	BBH (Acc.)	27.34	28.24	29.35	29.30	25.21	25.37	28.15	**29.44**	24.51	27.37	27.74	28.29
	Average	40.06	43.78	43.95	**45.68**	40.30	42.57	44.10	44.58	39.80	43.28	44.57	45.48
Mathematics	GSM8K (Pass@64)	57.59	66.86	66.89	73.64	59.40	70.02	75.25	**76.63**	56.16	66.38	72.31	74.60
	MATH-500 (Pass@64)	42.95	52.32	54.27	**57.82**	41.29	50.96	53.70	56.76	41.69	48.35	50.79	56.46
	Minerva (Pass@64)	16.68	18.55	20.30	21.75	17.23	18.24	21.93	**22.41**	17.37	19.38	19.95	21.71
	OlympiadBench (Pass@64)	22.26	24.79	24.54	24.58	20.53	22.27	22.77	24.31	22.04	25.70	24.86	**26.52**
	Average	34.87	40.63	41.50	44.45	34.61	40.37	43.41	**45.03**	34.32	39.95	41.98	44.82
Coding	HumanEval+ (Pass@64)	20.72	29.90	34.53	34.19	19.12	28.36	30.82	**34.66**	19.07	26.91	32.81	33.38
	MBPP+ (Pass@64)	52.05	65.95	70.15	**73.34**	51.01	63.64	66.74	70.16	51.40	62.94	68.06	72.47
	Average	36.39	47.93	52.34	**53.77**	35.07	46.00	48.78	52.41	35.24	44.93	50.44	52.93
Reasoning Average		37.11	44.11	45.93	**47.97**	36.66	42.98	45.43	47.34	36.45	42.72	45.66	47.74
Average		34.63	40.74	42.87	**44.89**	34.22	39.83	42.15	44.11	34.30	39.83	42.54	44.52

Table 6: **Pre-Training performance comparison across different $\tilde{\lambda}$ and $\hat{\lambda}$ based on 1B dense models. The highest scores at the final checkpoint across the different configurations are shown in bold.**

	Hyperparameters	$\beta = 0;\ \tilde{\lambda} = 0;\ \hat{\lambda} = -0.1$				$\beta = 0;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$				$\beta = 0;\ \tilde{\lambda} = 0.1;\ \hat{\lambda} = 0$			
	# Pre-Trained Tokens	125B	250B	375B	500B	125B	250B	375B	500B	125B	250B	375B	500B
General Knowledge	MMLU (Acc.)	21.61	23.71	24.04	26.06	20.85	23.39	24.04	**26.62**	23.59	22.92	24.63	25.89
	MMLU-Pro (Acc.)	8.34	9.32	8.42	8.67	7.63	8.23	7.98	9.02	9.04	8.69	8.50	**9.23**
	NaturalQuestions (EM)	2.66	4.71	4.90	5.21	2.74	4.07	4.52	**5.79**	3.05	4.79	4.74	5.15
	TriviaQA (EM)	8.61	12.70	14.95	15.59	8.65	12.63	13.53	**16.25**	8.76	11.18	14.08	15.50
	Average	10.31	12.61	13.08	13.88	9.97	12.08	12.52	**14.42**	11.11	11.90	12.99	13.94
Commonsense Reasoning	Hellaswag (Acc.)	38.61	44.06	46.40	48.17	38.96	44.42	46.26	**48.33**	38.87	43.72	46.67	48.26
	SIQA (Acc.)	39.51	38.84	40.63	40.48	40.02	40.33	41.50	42.32	38.79	40.53	41.30	**42.37**
	PIQA (Acc.)	66.70	69.64	70.84	70.95	67.90	69.75	71.00	**71.11**	67.30	69.48	71.60	**71.11**
	WinoGrande (Acc.)	50.20	50.36	51.54	52.33	52.49	52.17	52.80	**53.83**	48.93	51.38	54.78	52.64
	OpenBookQA (Acc.)	29.40	31.60	32.20	32.40	29.80	32.20	33.60	**32.80**	30.80	30.20	32.20	31.40
	CommonsenseQA (Acc.)	19.49	19.08	20.39	18.92	19.57	21.79	19.82	20.07	20.64	19.00	18.84	**23.34**
	Average	40.65	42.26	43.67	43.88	41.46	43.44	44.16	44.74	40.89	42.39	44.23	**44.85**
Knowledge Average		25.48	27.44	28.37	28.88	25.71	27.76	28.34	**29.58**	26.00	27.14	28.61	29.40
Logic Reasoning	ARC-e (Acc.)	52.36	57.28	59.09	58.46	51.64	55.68	56.70	58.80	50.21	55.01	58.38	**60.35**
	ARC-c (Acc.)	27.65	29.86	28.84	29.78	27.30	29.44	29.52	29.01	26.02	28.16	30.46	**30.20**
	BBH (Acc.)	24.48	24.33	25.19	23.42	22.13	22.56	25.68	**27.34**	23.32	24.42	24.05	25.59
	Average	34.83	37.16	37.71	37.22	33.69	35.89	37.30	38.38	33.18	35.86	37.63	**38.71**
Mathematics	GSM8K (Pass@64)	38.32	44.09	47.21	47.38	41.09	48.78	48.77	48.76	41.85	43.30	45.25	**50.04**
	MATH-500 (Pass@64)	34.17	36.42	38.42	**39.69**	33.41	35.43	37.10	38.49	33.52	38.01	37.35	37.29
	Minerva (Pass@64)	14.35	16.68	16.20	**17.07**	16.44	16.71	14.80	16.43	15.11	16.89	17.21	16.87
	OlympiadBench (Pass@64)	20.44	22.29	21.95	**23.76**	21.06	21.01	21.85	21.83	21.12	21.24	20.77	21.62
	Average	26.82	29.87	30.95	**31.98**	28.00	30.48	30.63	31.38	27.90	29.86	30.15	31.46
Coding	HumanEval+ (Pass@64)	8.13	13.04	15.12	14.69	8.71	12.15	12.72	15.95	8.84	12.34	16.91	**16.39**
	MBPP+ (Pass@64)	19.08	30.08	38.37	**43.02**	20.63	34.37	37.53	41.67	19.25	32.66	40.90	41.62
	Average	13.61	21.56	26.75	28.86	14.67	23.26	25.13	28.81	14.05	22.50	28.91	**29.01**
Reasoning Average		25.09	29.53	31.80	32.68	25.45	29.88	31.02	32.86	25.04	29.41	32.23	**33.06**
Average		25.24	28.69	30.43	31.16	25.56	29.03	29.95	31.55	25.43	28.50	30.78	**31.59**

Table 7: **Pre-Training performance comparison across different $\tilde{\lambda}$ and $\hat{\lambda}$ based on 4B dense models. The highest scores at the final checkpoint across the different configurations are shown in bold.**

	Hyperparameters	$\beta = 0;\ \tilde{\lambda} = 0;\ \hat{\lambda} = -0.1$				$\beta = 0;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$				$\beta = 0;\ \tilde{\lambda} = 0.1;\ \hat{\lambda} = 0$			
	# Pre-Trained Tokens	125B	250B	375B	500B	125B	250B	375B	500B	125B	250B	375B	500B
General Knowledge	MMLU (Acc.)	23.79	30.57	33.71	34.80	25.76	29.91	33.61	**36.17**	23.02	31.66	33.80	35.43
	MMLU-Pro (Acc.)	9.57	10.37	11.50	11.83	9.25	10.90	11.71	**12.23**	8.44	11.37	11.87	12.20
	NaturalQuestions (EM)	6.15	9.92	11.41	11.77	5.26	9.34	10.86	**12.35**	6.65	9.75	10.89	12.22
	TriviaQA (EM)	17.19	26.53	31.13	33.59	17.91	25.94	27.80	**33.83**	16.93	25.99	30.57	33.29
	Average	14.18	19.35	21.94	23.00	14.89	19.13	21.00	**23.61**	13.41	19.59	21.78	23.32
Commonsense Reasoning	Hellaswag (Acc.)	50.56	57.22	60.25	62.27	50.22	57.00	60.50	**62.91**	50.21	58.09	61.30	62.09
	SIQA (Acc.)	42.17	42.48	44.11	**46.98**	42.37	43.35	43.65	44.73	41.40	44.52	45.39	45.29
	PIQA (Acc.)	71.27	74.37	74.16	75.19	72.42	74.70	75.35	**76.28**	71.16	73.88	73.56	76.12
	WinoGrande (Acc.)	54.62	56.43	59.04	58.88	53.67	55.80	56.67	58.72	53.91	57.54	58.25	**59.91**
	OpenBookQA (Acc.)	34.40	38.60	38.00	**37.80**	33.40	36.40	37.00	36.00	32.80	35.80	36.40	37.40
	CommonsenseQA (Acc.)	19.41	30.06	39.80	46.76	20.39	28.26	47.91	**52.91**	20.56	30.06	41.52	46.93
	Average	45.41	49.86	52.56	54.65	45.41	49.25	53.51	**55.26**	45.01	49.98	52.74	54.62
Knowledge Average		29.79	34.60	37.25	38.82	30.15	34.19	37.26	**39.44**	29.21	34.79	37.26	38.97
Logic Reasoning	ARC-e (Acc.)	59.51	65.32	68.31	**68.14**	60.27	63.05	67.97	67.55	60.19	67.05	68.48	67.63
	ARC-c (Acc.)	31.57	35.15	36.95	37.20	30.55	33.96	36.77	**37.54**	31.66	35.75	38.23	37.20
	BBH (Acc.)	27.38	28.08	**29.15**	29.15	26.54	27.42	29.53	28.29	26.22	27.15	30.32	28.80
	Average	39.49	42.85	44.80	**44.83**	39.12	41.48	44.76	44.46	39.36	43.32	45.68	44.54
Mathematics	GSM8K (Pass@64)	46.40	59.61	62.93	70.92	48.43	60.24	68.75	**71.26**	48.09	58.17	61.84	70.94
	MATH-500 (Pass@64)	38.86	47.09	48.23	**52.47**	38.19	48.15	48.88	51.54	38.16	43.83	46.77	50.45
	Minerva (Pass@64)	16.60	17.99	17.90	**21.27**	15.09	19.22	19.68	20.63	16.67	18.06	19.47	20.54
	OlympiadBench (Pass@64)	21.80	23.30	24.40	24.19	21.79	25.53	23.70	24.84	22.92	25.74	25.16	**25.62**
	Average	30.92	37.00	38.37	**42.21**	30.88	38.29	40.25	42.07	31.46	36.45	38.31	41.89
Coding	HumanEval+ (Pass@64)	16.71	23.77	27.27	29.41	17.32	23.22	30.08	29.13	15.42	20.41	29.56	**31.08**
	MBPP+ (Pass@64)	42.24	56.78	63.88	66.01	40.66	56.64	65.69	**66.29**	38.19	57.32	64.74	65.65
	Average	29.48	40.28	45.58	47.71	28.99	39.93	47.89	47.71	26.81	38.87	47.15	**48.37**
Reasoning Average		33.29	40.04	42.91	44.92	33.00	39.90	44.30	44.75	32.54	39.54	43.71	**44.93**
Average		31.89	37.87	40.65	42.48	31.86	37.61	41.48	**42.62**	29.21	37.64	41.13	42.55

Table 8: **Pre-Training performance comparison across different $\tilde{\lambda}$ and $\hat{\lambda}$ based on 5B-A0.3B MoE models. The highest scores at the final checkpoint across the different configurations are shown in bold.**

	Hyperparameters	$\beta = 0;\ \tilde{\lambda} = 0;\ \hat{\lambda} = -0.1$				$\beta = 0;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$				$\beta = 0;\ \tilde{\lambda} = 0.1;\ \hat{\lambda} = 0$			
	# Pre-Trained Tokens	125B	250B	375B	500B	125B	250B	375B	500B	125B	250B	375B	500B
General Knowledge	MMLU (Acc.)	22.24	26.11	26.41	29.99	23.10	25.12	29.17	31.19	22.20	26.45	29.44	**31.37**
	MMLU-Pro (Acc.)	7.55	8.74	8.82	10.36	9.47	8.89	10.06	**11.87**	8.18	9.63	10.11	11.02
	NaturalQuestions (EM)	5.60	7.34	7.37	8.84	5.15	7.51	8.73	**10.42**	5.21	7.95	8.31	9.81
	TriviaQA (EM)	14.54	21.20	25.70	27.26	14.93	22.03	25.27	28.25	15.09	21.84	25.85	**28.91**
	Average	12.48	15.85	17.08	19.11	13.16	15.89	18.31	**20.43**	12.67	16.47	18.43	20.28
Commonsense Reasoning	Hellaswag (Acc.)	48.44	53.64	56.72	**57.58**	48.54	54.63	56.88	57.41	48.51	53.89	56.42	57.33
	SIQA (Acc.)	40.48	43.55	43.65	**42.84**	41.25	42.27	43.76	42.68	38.84	41.50	41.50	41.81
	PIQA (Acc.)	69.48	72.96	73.39	74.37	71.27	73.18	74.43	74.76	71.55	74.43	74.43	**74.92**
	WinoGrande (Acc.)	52.80	54.85	55.01	55.56	50.43	53.20	56.51	**56.20**	52.88	54.22	56.91	55.96
	OpenBookQA (Acc.)	32.00	31.80	33.40	34.20	31.20	32.80	34.40	33.60	32.20	34.40	34.40	**36.40**
	CommonsenseQA (Acc.)	19.74	23.67	20.15	30.30	18.67	22.93	28.50	**35.38**	20.07	24.65	27.76	31.86
	Average	43.82	46.75	47.05	49.14	43.56	46.50	49.08	**50.01**	44.01	47.18	48.57	49.71
Knowledge Average		28.15	31.30	32.06	34.13	28.36	31.19	33.69	**35.22**	28.34	31.82	33.50	35.00
Logic Reasoning	ARC-e (Acc.)	57.15	61.49	63.43	63.01	58.54	62.92	64.48	**63.34**	58.63	62.25	62.42	62.08
	ARC-c (Acc.)	30.20	33.28	34.73	33.62	30.80	34.81	35.67	**35.15**	31.06	34.73	34.04	34.98
	BBH (Acc.)	24.56	24.60	28.29	26.45	23.97	25.79	27.71	27.03	22.33	26.28	28.09	**27.74**
	Average	37.30	39.79	42.15	41.03	37.77	41.17	42.62	**41.84**	37.34	41.09	41.52	41.60
Mathematics	GSM8K (Pass@64)	50.99	59.44	62.10	64.61	51.48	59.15	61.81	**65.65**	48.16	58.15	60.69	59.90
	MATH-500 (Pass@64)	39.91	46.65	48.92	50.65	37.57	43.69	46.91	**51.19**	38.71	42.31	46.49	48.60
	Minerva (Pass@64)	16.36	15.98	18.30	**18.83**	15.47	16.93	18.00	17.92	17.31	18.01	17.78	18.02
	OlympiadBench (Pass@64)	22.16	23.01	25.23	**25.54**	23.50	24.14	23.40	24.42	22.62	24.02	24.57	24.19
	Average	32.36	36.27	38.64	**39.91**	32.01	35.98	37.53	39.80	31.70	35.62	37.38	37.68
Coding	HumanEval+ (Pass@64)	14.01	22.49	27.53	27.00	17.52	22.79	26.15	**29.32**	15.27	21.36	24.62	27.88
	MBPP+ (Pass@64)	38.32	52.88	57.44	**60.50**	39.81	54.29	55.93	57.77	40.39	49.00	56.04	60.17
	Average	26.17	37.69	42.49	43.75	28.67	38.54	41.04	43.55	27.83	35.18	40.33	**44.03**
Reasoning Average		31.94	37.92	41.09	41.56	32.81	38.56	40.40	**41.73**	32.29	37.30	39.74	41.10
Average		30.43	35.27	37.48	38.59	31.03	35.62	37.72	**39.12**	30.71	35.11	37.25	38.66

Table 9: **Pre-Training performance comparison across different $\tilde{\lambda}$ and $\hat{\lambda}$ based on 10B-A0.5B MoE models. The highest scores at the final checkpoint across the different configurations are shown in bold.**

	Hyperparameters	$\beta = 0;\ \tilde{\lambda} = 0;\ \hat{\lambda} = -0.1$				$\beta = 0;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$				$\beta = 0;\ \tilde{\lambda} = 0.1;\ \hat{\lambda} = 0$			
	# Pre-Trained Tokens	125B	250B	375B	500B	125B	250B	375B	500B	125B	250B	375B	500B
General Knowledge	MMLU (Acc.)	25.99	31.87	35.24	36.13	25.35	29.14	33.50	35.12	25.38	32.48	34.47	**36.75**
	MMLU-Pro (Acc.)	9.40	11.10	12.17	**13.19**	8.24	10.01	11.64	12.15	8.96	11.19	12.64	12.34
	NaturalQuestions (EM)	7.51	11.22	11.94	12.60	7.06	10.47	11.41	11.96	7.06	10.44	11.25	**13.52**
	TriviaQA (EM)	20.55	29.22	33.57	37.30	19.94	29.86	33.89	37.31	21.12	29.48	33.93	**37.61**
	Average	15.86	20.85	23.23	24.81	15.15	19.87	22.61	24.14	15.63	20.90	23.07	**25.06**
Commonsense Reasoning	Hellaswag (Acc.)	52.70	59.23	63.30	63.68	53.00	59.27	62.00	**63.71**	53.06	59.75	63.67	63.38
	SIQA (Acc.)	40.63	43.60	44.37	44.17	42.02	42.73	43.45	45.09	41.04	44.17	45.34	**45.50**
	PIQA (Acc.)	73.23	74.92	75.90	76.22	72.74	74.59	75.52	**76.71**	72.42	74.70	75.52	76.50
	WinoGrande (Acc.)	52.72	57.54	58.56	**59.98**	52.88	56.12	59.04	58.88	53.83	57.14	59.12	**59.98**
	OpenBookQA (Acc.)	34.20	35.60	37.00	36.80	32.80	35.60	35.60	**38.20**	33.60	35.40	36.20	35.60
	CommonsenseQA (Acc.)	21.38	37.92	44.64	**48.89**	22.52	33.74	35.54	43.90	19.25	34.81	41.03	45.29
	Average	45.81	51.47	53.96	**54.96**	45.99	50.34	51.86	54.42	45.53	51.00	53.48	54.38
Knowledge Average		30.84	36.16	38.60	**39.88**	30.57	35.11	37.23	39.28	30.58	35.95	38.28	39.72
Logic Reasoning	ARC-e (Acc.)	60.69	66.25	69.61	**68.43**	63.01	67.09	67.63	67.00	64.06	65.99	70.03	67.59
	ARC-c (Acc.)	33.87	37.63	38.73	37.71	32.68	35.24	36.52	37.29	34.13	37.71	37.88	**39.42**
	BBH (Acc.)	25.63	26.34	29.06	29.10	25.21	25.37	28.15	29.44	25.20	29.46	29.72	**29.70**
	Average	40.06	43.41	45.80	45.08	40.30	42.57	44.10	44.58	41.13	44.39	45.88	**45.57**
Mathematics	GSM8K (Pass@64)	55.13	69.42	71.80	**77.63**	59.40	70.02	75.25	76.63	55.89	65.29	71.44	75.11
	MATH-500 (Pass@64)	43.56	49.61	54.32	56.39	41.29	50.96	53.70	**56.76**	42.58	49.50	52.46	54.75
	Minerva (Pass@64)	17.23	19.23	20.27	**22.60**	17.23	18.24	21.93	22.41	16.07	19.55	21.59	20.38
	OlympiadBench (Pass@64)	22.90	22.84	24.30	**26.67**	20.53	22.27	22.77	24.31	23.23	24.28	25.25	26.15
	Average	34.71	40.28	42.67	**45.82**	34.61	40.37	43.41	45.03	34.44	39.66	42.69	44.10
Coding	HumanEval+ (Pass@64)	21.82	29.67	32.92	**35.18**	19.12	28.36	30.82	34.66	19.54	25.32	30.97	32.96
	MBPP+ (Pass@64)	48.31	63.41	68.98	71.01	51.01	63.64	66.74	70.16	50.24	60.62	68.22	**71.02**
	Average	35.07	46.54	50.95	**53.10**	35.07	46.00	48.78	52.41	34.89	42.97	49.60	51.99
Reasoning Average		36.61	43.41	46.47	**48.00**	36.66	42.98	45.43	47.34	36.82	42.34	46.05	47.22
Average		34.30	40.51	43.32	**44.75**	34.22	39.83	42.15	44.11	34.33	39.78	42.94	44.22

Table 10: **Mid-Training performance comparison across different β based on 4B dense models. The highest scores at the final checkpoint across the different configurations are shown in bold.**

	Hyperparameters	$\beta = -0.25;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$				$\beta = 0;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$				$\beta = 0.50;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$			
	# Mid-Trained Tokens	25B	50B	75B	100B	25B	50B	75B	100B	25B	50B	75B	100B
General Knowledge	MMLU (Acc.)	37.99	39.51	39.69	39.92	37.27	39.34	39.23	**40.47**	36.53	38.45	39.00	39.48
	MMLU-Pro (Acc.)	16.51	18.44	19.69	**19.70**	15.50	17.27	17.77	18.68	14.97	16.44	17.99	18.53
	NaturalQuestions (EM)	11.69	11.75	12.16	12.33	12.16	11.91	12.33	**12.55**	11.75	11.14	12.08	11.88
	TriviaQA (EM)	32.60	32.75	34.05	34.50	31.90	32.54	33.50	34.17	32.55	32.78	34.09	**34.67**
	Average	24.70	25.61	26.40	**26.61**	24.21	25.27	25.71	26.47	23.95	24.70	25.79	26.14
Commonsense Reasoning	Hellaswag (Acc.)	61.59	61.70	61.95	62.41	61.25	61.72	62.14	62.09	61.68	62.01	62.54	**62.51**
	SIQA (Acc.)	44.78	43.71	44.42	44.78	45.44	45.29	45.75	**45.85**	44.32	44.78	44.83	44.88
	PIQA (Acc.)	75.24	76.01	75.19	**75.84**	75.24	75.24	75.24	75.73	74.54	74.92	74.81	74.92
	WinoGrande (Acc.)	60.69	59.67	60.22	60.62	58.96	59.98	60.46	**60.85**	58.88	59.35	59.51	59.67
	OpenBookQA (Acc.)	35.60	38.20	37.20	37.40	36.60	37.40	38.00	37.40	39.20	40.40	40.60	**40.00**
	CommonsenseQA (Acc.)	52.09	52.42	54.38	**55.77**	54.13	52.99	53.97	54.87	49.14	50.04	52.58	53.07
	Average	55.00	55.29	55.56	**56.14**	55.27	55.44	55.93	56.13	54.63	55.25	55.81	55.84
Knowledge Average		39.85	40.45	40.98	**41.37**	39.74	40.35	40.82	41.30	39.29	39.98	40.80	40.99
Logic Reasoning	ARC-e (Acc.)	66.84	68.94	70.45	69.82	68.81	68.94	69.82	69.99	66.84	69.19	69.91	**71.17**
	ARC-c (Acc.)	39.33	41.04	40.78	**41.89**	40.61	41.13	41.64	**41.89**	38.14	41.38	40.87	41.72
	BBH (Acc.)	33.90	38.58	39.66	**39.83**	33.47	36.95	36.45	37.28	31.90	34.88	36.11	35.92
	Average	46.69	49.52	50.30	**50.51**	47.63	49.01	49.30	49.72	45.63	48.48	48.96	49.60
Mathematics	GSM8K (Pass@64)	84.66	88.94	91.17	91.25	85.73	89.60	92.23	**92.70**	83.30	88.82	91.75	92.08
	MATH-500 (Pass@64)	63.69	66.70	70.73	70.41	62.42	66.14	69.48	**71.35**	62.42	65.49	68.17	68.97
	Minerva (Pass@64)	21.88	23.67	26.82	25.99	24.70	24.38	28.22	**26.72**	22.86	23.90	23.95	24.99
	OlympiadBench (Pass@64)	28.96	30.15	32.15	32.21	28.44	33.15	32.96	32.65	29.21	29.80	31.51	**32.85**
	Average	49.80	52.37	55.22	54.97	50.32	53.32	55.72	**55.86**	49.45	52.00	53.85	54.72
Coding	HumanEval+ (Pass@64)	52.37	57.74	64.85	64.33	51.64	60.52	64.95	65.89	48.07	60.60	63.31	**66.30**
	MBPP+ (Pass@64)	79.91	82.27	85.48	**86.79**	77.50	82.32	84.81	85.28	77.83	82.51	83.77	83.41
	Average	66.74	70.01	75.17	75.56	64.57	71.42	74.88	**75.59**	62.95	71.56	73.54	74.86
Reasoning Average		54.21	57.30	60.23	60.35	54.17	57.91	59.97	**60.39**	52.67	57.35	58.78	59.73
Average		48.46	50.56	52.53	**52.76**	48.40	50.89	52.31	52.75	47.32	50.40	51.59	52.23

Table 11: **Mid-Training performance comparison across different β based on 10B-A0.5B MoE models. The highest scores at the final checkpoint across the different configurations are shown in bold.**

	Hyperparameters	$\beta = -0.25;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$				$\beta = 0;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$				$\beta = 0.50;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$			
	# Mid-Trained Tokens	25B	50B	75B	100B	25B	50B	75B	100B	25B	50B	75B	100B
General Knowledge	MMLU (Acc.)	36.39	37.23	37.75	**38.11**	36.15	37.27	37.54	37.82	35.04	36.49	36.45	37.10
	MMLU-Pro (Acc.)	12.31	12.88	13.69	13.91	12.98	13.91	14.41	**14.71**	12.03	12.82	13.32	13.46
	NaturalQuestions (EM)	12.80	13.77	14.18	**14.35**	11.30	12.11	12.44	12.85	11.30	12.11	12.44	12.85
	TriviaQA (EM)	36.61	37.37	38.29	38.76	35.84	37.15	37.81	38.38	37.57	38.84	39.62	**40.21**
	Average	24.53	25.31	25.98	**26.28**	24.07	25.11	25.55	25.94	24.40	25.41	25.87	26.18
Commonsense Reasoning	Hellaswag (Acc.)	62.25	62.39	62.94	63.25	62.47	62.82	63.50	**63.74**	62.51	62.67	63.29	63.43
	SIQA (Acc.)	43.86	44.63	44.68	**45.19**	45.55	43.04	43.70	44.22	44.63	44.01	44.73	44.01
	PIQA (Acc.)	75.46	74.92	75.57	75.95	76.06	75.63	75.73	76.55	75.90	76.06	76.50	**76.82**
	WinoGrande (Acc.)	58.80	59.98	60.54	**60.46**	58.96	58.88	59.59	59.83	58.64	59.51	60.77	60.30
	OpenBookQA (Acc.)	37.40	36.80	37.60	38.00	36.60	37.20	37.00	37.20	40.00	40.40	40.20	**40.60**
	CommonsenseQA (Acc.)	50.53	50.45	50.61	**50.61**	43.90	46.52	49.06	49.06	40.46	42.75	43.73	43.90
	Average	54.72	54.86	55.32	**55.58**	53.92	54.02	54.76	55.10	53.69	54.23	54.87	54.84
Knowledge Average		39.62	40.09	40.65	**40.93**	39.00	39.56	40.16	40.52	39.04	39.82	40.37	40.51
Logic Reasoning	ARC-e (Acc.)	68.35	68.90	69.40	70.12	68.35	68.90	69.40	70.12	68.52	70.03	69.87	**70.24**
	ARC-c (Acc.)	38.65	40.10	41.30	40.53	39.59	40.44	42.49	**42.83**	39.93	40.87	41.64	41.98
	BBH (Acc.)	30.70	31.64	31.84	32.05	30.64	33.68	33.59	**34.14**	30.69	32.42	33.25	33.44
	Average	45.90	46.88	47.51	47.57	46.40	48.30	48.91	**49.27**	46.38	47.77	48.25	48.55
Mathematics	GSM8K (Pass@64)	85.06	87.67	90.35	90.53	87.16	90.68	90.83	**92.03**	84.62	88.39	89.84	90.30
	MATH-500 (Pass@64)	65.20	68.85	70.19	**71.73**	64.39	68.35	70.77	70.97	64.27	68.59	67.46	70.00
	Minerva (Pass@64)	25.33	25.02	26.26	26.53	24.51	23.83	24.77	**27.03**	23.51	26.76	27.11	25.90
	OlympiadBench (Pass@64)	29.73	30.76	32.84	33.21	28.91	30.05	33.11	32.63	28.57	29.86	32.82	**33.39**
	Average	51.33	53.08	54.91	55.50	51.24	53.23	54.87	**55.67**	50.24	53.40	54.31	54.90
Coding	HumanEval+ (Pass@64)	48.79	55.88	56.24	**58.05**	48.00	53.35	56.98	55.18	49.33	53.14	55.06	56.21
	MBPP+ (Pass@64)	77.31	80.74	84.55	**83.41**	76.28	80.29	81.82	81.38	76.03	78.12	80.15	80.80
	Average	63.05	68.31	70.40	**70.73**	62.14	66.82	69.40	68.28	62.68	65.63	67.61	68.51
Reasoning Average		53.43	56.09	57.61	**57.93**	53.26	56.12	57.73	57.74	53.10	55.60	56.72	57.32
Average		47.90	49.69	50.82	**51.13**	47.56	49.50	50.70	50.85	47.48	49.29	50.18	50.60

Table 12: **Mid-Training performance comparison across different $\tilde{\lambda}$ and $\hat{\lambda}$ based on 4B dense models. The highest scores at the final checkpoint across the different configurations are shown in bold.**

	Hyperparameters	$\beta = 0;\ \tilde{\lambda} = 0;\ \hat{\lambda} = -0.1$				$\beta = 0;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$				$\beta = 0;\ \tilde{\lambda} = 0.1;\ \hat{\lambda} = 0$			
	# Mid-Trained Tokens	25B	50B	75B	100B	25B	50B	75B	100B	25B	50B	75B	100B
General Knowledge	MMLU (Acc.)	37.33	38.73	38.83	39.13	37.27	39.34	39.23	**40.47**	37.15	38.06	38.76	39.43
	MMLU-Pro (Acc.)	14.17	16.61	17.57	18.23	15.50	17.27	17.77	**18.68**	14.17	15.58	16.41	17.50
	NaturalQuestions (EM)	11.25	11.99	12.30	12.66	12.16	11.91	12.33	12.55	11.14	12.49	12.58	**12.74**
	TriviaQA (EM)	33.29	33.89	34.80	**35.32**	31.90	32.54	33.50	34.17	32.73	33.50	34.17	34.59
	Average	24.01	25.31	25.88	26.34	24.21	25.27	25.71	**26.47**	23.80	24.91	25.48	26.07
Commonsense Reasoning	Hellaswag (Acc.)	61.11	61.25	61.76	**62.29**	61.25	61.72	62.14	62.09	61.59	61.46	61.98	62.16
	SIQA (Acc.)	45.09	44.37	44.06	45.24	45.44	45.29	45.75	45.85	46.21	46.21	46.37	**46.62**
	PIQA (Acc.)	75.46	76.01	76.22	**76.12**	75.24	75.24	75.24	75.73	74.59	74.76	75.41	75.41
	WinoGrande (Acc.)	60.22	60.22	61.48	60.46	58.96	59.98	60.46	**60.85**	60.85	60.14	59.51	59.98
	OpenBookQA (Acc.)	39.80	38.20	38.20	38.00	36.60	37.40	38.00	37.40	39.80	38.60	39.60	**40.20**
	CommonsenseQA (Acc.)	46.93	51.84	53.73	**53.81**	54.13	52.99	53.97	54.87	45.86	47.67	49.80	50.20
	Average	54.77	55.32	55.91	55.99	55.27	55.44	55.93	**56.13**	54.82	54.81	55.45	55.76
Knowledge Average		39.39	40.31	40.89	41.16	39.74	40.35	40.82	**41.30**	39.31	39.86	40.46	40.91
Logic Reasoning	ARC-e (Acc.)	68.52	68.56	70.75	**71.34**	68.81	68.94	69.82	69.99	70.33	70.29	71.04	71.25
	ARC-c (Acc.)	39.93	40.96	41.81	42.66	40.61	41.13	41.64	41.89	41.30	42.15	42.24	**43.34**
	BBH (Acc.)	32.78	37.15	38.49	**39.30**	33.47	36.95	36.45	37.28	33.62	35.40	36.58	37.74
	Average	47.08	48.89	50.35	**51.10**	47.63	49.01	49.30	49.72	48.42	49.28	49.95	50.78
Mathematics	GSM8K (Pass@64)	84.38	89.90	90.43	91.69	85.73	89.60	92.23	**92.70**	85.99	87.51	90.68	90.11
	MATH-500 (Pass@64)	64.40	70.24	72.50	**72.78**	62.42	66.14	69.48	71.35	63.31	67.35	70.36	70.31
	Minerva (Pass@64)	21.64	23.81	25.65	**26.86**	24.70	24.38	28.22	26.72	23.28	22.60	25.53	25.43
	OlympiadBench (Pass@64)	30.09	32.29	33.52	**34.13**	28.44	33.15	32.96	32.65	26.32	28.90	30.55	34.11
	Average	50.13	54.06	55.53	**56.37**	50.32	53.32	55.72	55.86	49.73	51.59	54.28	54.99
Coding	HumanEval+ (Pass@64)	50.62	58.51	64.58	65.23	51.64	60.52	64.95	**65.89**	51.68	58.17	63.94	65.31
	MBPP+ (Pass@64)	78.24	83.29	85.36	85.22	77.50	82.32	84.81	**85.28**	79.31	81.59	85.17	85.23
	Average	64.43	70.90	74.97	75.23	64.57	71.42	74.88	**75.59**	65.50	69.88	74.56	75.27
Reasoning Average		53.88	57.95	60.28	**60.90**	54.17	57.91	59.97	60.39	54.55	56.92	59.60	60.35
Average		48.08	50.89	52.53	**53.00**	48.40	50.89	52.31	52.75	48.45	50.09	51.94	52.57

Table 13: **Mid-Training performance comparison across different $\tilde{\lambda}$ and $\hat{\lambda}$ based on 10B-A0.5B MoE models. The highest scores at the final checkpoint across the different configurations are shown in bold.**

	Hyperparameters	$\beta = 0;\ \tilde{\lambda} = 0;\ \hat{\lambda} = -0.1$				$\beta = 0;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$				$\beta = 0;\ \tilde{\lambda} = 0.1;\ \hat{\lambda} = 0$			
	# Mid-Trained Tokens	25B	50B	75B	100B	25B	50B	75B	100B	25B	50B	75B	100B
General Knowledge	MMLU (Acc.)	36.86	38.22	38.98	39.07	36.15	37.27	37.54	37.82	36.90	37.96	38.29	**39.30**
	MMLU-Pro (Acc.)	14.10	15.20	15.93	**16.39**	12.98	13.91	14.41	14.71	13.02	15.05	14.67	14.72
	NaturalQuestions (EM)	12.35	13.05	13.63	**13.60**	11.30	12.11	12.44	12.85	12.52	12.63	13.30	13.49
	TriviaQA (EM)	35.51	37.14	37.49	38.25	35.84	37.15	37.81	38.38	36.25	36.98	38.01	**38.44**
	Average	24.71	25.90	26.51	**26.83**	24.07	25.11	25.55	25.94	24.67	25.66	26.07	26.49
Commonsense Reasoning	Hellaswag (Acc.)	62.67	62.96	63.12	63.50	62.47	62.82	63.50	**63.74**	62.93	63.07	63.42	63.73
	SIQA (Acc.)	43.86	43.96	43.71	43.71	45.55	43.04	43.70	44.22	45.34	45.50	45.91	**45.45**
	PIQA (Acc.)	76.01	75.41	75.95	76.22	76.06	75.63	75.73	**76.55**	76.01	75.73	76.17	76.12
	WinoGrande (Acc.)	60.06	61.09	60.77	58.88	60.06	61.09	60.77	58.88	58.56	59.59	60.46	**61.01**
	OpenBookQA (Acc.)	37.60	35.80	36.00	35.40	36.60	37.20	37.00	37.20	37.20	35.80	37.00	**37.80**
	CommonsenseQA (Acc.)	45.62	48.89	45.86	48.48	43.90	46.52	49.06	49.06	45.45	47.83	47.91	**49.63**
	Average	54.30	54.69	54.24	54.37	53.92	54.02	54.76	55.10	54.25	54.59	55.15	**55.62**
Knowledge Average		39.50	40.29	40.37	40.60	39.00	39.56	40.16	40.52	39.46	40.12	40.61	**41.06**
Logic Reasoning	ARC-e (Acc.)	69.78	69.78	70.58	71.09	68.98	70.79	70.66	70.83	68.48	69.95	70.56	**71.13**
	ARC-c (Acc.)	40.10	41.21	41.81	41.30	39.59	40.44	42.49	**42.83**	41.64	42.41	41.81	42.06
	BBH (Acc.)	31.18	33.21	33.71	34.05	30.64	33.68	33.59	**34.14**	30.70	33.19	32.64	32.73
	Average	47.02	48.07	48.70	48.81	46.40	48.30	48.91	**49.27**	46.94	48.52	48.34	48.64
Mathematics	GSM8K (Pass@64)	87.23	90.34	91.10	91.45	87.16	90.68	90.83	**92.03**	84.20	87.47	89.89	90.41
	MATH-500 (Pass@64)	64.74	69.10	71.01	70.77	64.39	68.35	70.77	**70.97**	63.07	65.99	69.52	69.96
	Minerva (Pass@64)	24.50	25.23	25.50	26.67	24.51	23.83	24.77	27.03	23.05	26.98	27.02	**27.41**
	OlympiadBench (Pass@64)	28.84	30.41	31.34	31.20	28.91	30.05	33.11	32.63	31.11	32.09	32.57	**32.83**
	Average	51.33	53.77	54.74	55.02	51.24	53.23	54.87	**55.67**	50.36	53.13	54.75	55.15
Coding	HumanEval+ (Pass@64)	47.08	50.89	52.11	55.14	48.00	53.35	56.98	55.18	47.36	51.56	55.66	**55.62**
	MBPP+ (Pass@64)	78.67	79.80	81.25	**84.00**	76.28	80.29	81.82	81.38	76.24	80.35	82.42	83.16
	Average	62.88	65.35	66.68	**69.57**	62.14	66.82	69.40	68.28	61.80	65.96	69.04	69.39
Reasoning Average		53.74	55.73	56.71	**57.80**	53.26	56.12	57.73	57.74	53.03	55.87	57.38	57.73
Average		48.05	49.55	50.17	50.92	47.56	49.50	50.70	50.85	47.60	49.57	50.67	**51.06**

Table 14: **RL performance of the 4B Dense Model with $\beta = 0$; $\tilde{\lambda} = 0$; $\hat{\lambda} = 0$.**

	# RL Steps	100	200	300	400	500	600	700	800	900	1000
AIME24	Avg@128	0.89	0.81	1.14	1.04	1.12	1.54	1.33	1.69	1.64	1.62
	Cons@128	0.00	0.00	0.00	0.00	3.33	3.33	3.33	6.67	3.33	3.33
	Pass@64	15.34	13.77	21.27	14.80	11.68	22.86	19.81	19.68	14.78	18.32
AIME25	Avg@128	0.57	0.86	0.86	1.46	1.59	2.29	1.30	2.06	2.21	2.24
	Cons@128	0.00	0.00	0.00	3.33	0.00	6.67	3.33	3.33	3.33	3.33
	Pass@64	14.90	11.67	10.84	10.00	16.16	10.84	12.93	14.35	10.84	14.59
AMC23	Avg@128	18.46	23.07	24.94	27.87	28.96	28.73	27.17	28.18	29.16	29.51
	Cons@128	37.50	35.00	40.00	47.50	45.00	45.00	45.00	45.00	47.50	45.00
	Pass@64	71.39	76.75	79.44	75.70	74.24	72.45	67.17	77.07	78.26	77.43
OlympiadBench	Avg@128	14.53	17.94	19.98	21.45	22.31	22.28	22.04	23.67	24.64	24.56
	Cons@128	23.41	26.22	27.11	28.00	28.44	29.63	29.48	32.00	33.04	33.78
	Pass@64	55.45	55.01	56.70	56.49	55.80	55.84	55.45	56.73	57.57	58.09
MATH-500	Avg@128	39.85	45.48	48.23	50.56	51.41	51.10	49.87	51.22	52.20	52.26
	Cons@128	56.80	60.60	62.60	62.00	61.60	62.40	61.80	62.60	65.40	64.60
	Pass@64	87.63	87.16	87.38	88.36	88.88	88.81	88.82	89.49	90.18	89.95
Minerva	Avg@128	9.11	10.52	11.19	11.99	12.52	12.36	12.93	13.26	12.43	12.75
	Cons@128	17.65	17.28	18.38	17.64	19.12	19.49	18.01	20.59	19.12	19.49
	Pass@64	44.54	45.07	47.02	44.85	43.97	45.29	48.84	48.53	47.36	47.55
Average	Avg@128	13.90	16.45	17.72	19.06	19.65	19.72	19.11	20.01	20.38	20.49
	Cons@128	22.56	23.18	24.68	26.41	26.25	27.75	26.83	28.37	28.62	28.26
	Pass@64	48.21	48.24	50.44	48.37	48.46	49.35	48.84	50.98	49.83	50.99

Table 15: **RL performance of the 10B-A0.5B MoE Model with $\beta = 0$; $\tilde{\lambda} = 0$; $\hat{\lambda} = 0$.**

	# RL Steps	100	200	300	400	500	600	700	800	900	1000
AIME24	Avg@128	0.34	0.47	0.34	0.44	0.63	0.57	0.55	0.55	0.65	0.65
	Cons@128	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.33	0.00	0.00
	Pass@64	12.88	12.94	11.42	11.68	12.93	11.47	9.59	9.17	10.00	10.84
AIME25	Avg@128	0.13	0.23	0.16	0.29	0.36	0.47	0.60	0.44	0.63	0.57
	Cons@128	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Pass@64	6.68	11.27	8.35	14.80	16.58	19.10	18.67	14.10	16.66	16.59
AMC23	Avg@128	10.64	11.82	11.95	13.07	14.26	16.11	16.21	17.79	18.38	19.06
	Cons@128	20.00	25.00	17.50	27.50	25.00	30.00	27.50	35.00	32.50	37.50
	Pass@64	75.83	76.02	70.67	71.55	73.75	74.33	73.13	73.22	71.65	73.71
OlympiadBench	Avg@128	7.13	8.32	9.22	10.66	11.83	12.45	12.60	13.50	14.41	15.03
	Cons@128	14.81	16.00	17.04	20.00	21.04	21.33	21.19	22.52	24.30	24.89
	Pass@64	49.58	49.53	52.16	52.06	52.69	52.50	53.48	53.34	52.67	53.31
MATH-500	Avg@128	26.71	28.91	32.72	34.81	36.45	37.67	38.47	39.60	41.38	42.19
	Cons@128	45.60	48.40	49.80	52.40	54.40	54.40	55.20	56.20	58.20	58.20
	Pass@64	82.96	82.71	83.72	83.39	85.93	85.79	85.64	86.01	86.04	85.75
Minerva	Avg@128	5.66	5.70	6.42	6.93	7.22	7.84	7.98	8.28	8.36	8.73
	Cons@128	9.93	10.66	11.40	12.50	11.40	13.24	12.87	12.50	12.50	13.97
	Pass@64	42.39	40.93	41.04	42.75	43.47	42.89	44.35	44.50	40.90	44.90
Average	Avg@128	8.44	9.24	10.14	11.03	11.79	12.52	12.74	13.36	13.97	14.37
	Cons@128	15.06	16.68	15.96	18.73	18.64	19.83	19.46	21.59	21.25	22.43
	Pass@64	45.05	45.57	44.56	46.04	47.56	47.68	47.48	46.72	46.32	47.52

Table 16: **RL performance of the 4B Dense Model with $\beta = -0.25$; $\tilde{\lambda} = 0$; $\hat{\lambda} = 0$.**

	# RL Steps	100	200	300	400	500	600	700	800	900	1000
AIME24	Avg@128	0.76	0.96	1.02	0.55	1.09	1.43	1.82	2.03	2.47	2.57
	Cons@128	0.00	3.33	3.33	0.00	0.00	3.33	3.33	6.67	6.67	6.67
	Pass@64	14.17	17.11	13.31	12.10	17.09	19.99	20.64	19.17	24.12	20.34
AIME25	Avg@128	0.83	1.46	2.11	1.72	2.27	2.11	1.98	2.19	2.40	2.63
	Cons@128	0.00	0.00	3.33	3.33	3.33	3.33	3.33	3.33	6.67	3.33
	Pass@64	17.46	13.33	21.90	19.12	19.99	15.64	14.15	16.51	11.67	15.82
AMC23	Avg@128	19.55	25.47	29.08	30.96	31.70	31.72	32.38	32.85	33.01	33.09
	Cons@128	32.50	40.00	40.00	45.00	45.00	50.00	50.00	45.00	50.00	50.00
	Pass@64	72.21	79.29	75.65	74.33	72.62	72.48	70.62	71.24	76.36	74.00
OlympiadBench	Avg@128	14.44	17.77	19.71	20.38	21.52	21.44	22.98	23.33	23.52	23.67
	Cons@128	24.89	26.52	28.44	28.15	28.59	28.89	31.26	30.81	31.11	31.26
	Pass@64	54.46	56.71	57.38	55.02	58.18	55.17	57.97	57.45	56.56	57.23
MATH-500	Avg@128	39.36	45.23	48.29	49.88	50.59	50.72	52.33	52.50	53.09	52.98
	Cons@128	58.00	60.40	62.00	60.80	63.20	62.40	62.20	61.60	61.80	62.20
	Pass@64	87.37	87.68	88.34	88.53	88.89	89.18	89.02	88.82	89.21	89.94
Minerva	Avg@128	9.23	10.71	11.16	12.03	12.14	12.22	12.97	12.73	12.99	13.11
	Cons@128	18.39	18.38	18.75	17.65	18.75	18.38	20.96	18.75	18.38	19.49
	Pass@64	43.57	46.44	46.34	46.35	46.30	45.61	47.17	45.57	46.61	45.22
Average	Avg@128	14.03	16.93	18.56	19.25	19.89	19.94	20.74	20.94	21.25	21.34
	Cons@128	22.30	24.77	25.98	25.82	26.48	27.72	28.51	27.69	29.11	28.83
	Pass@64	48.21	50.09	50.49	49.24	50.51	49.68	49.93	49.79	50.76	50.43

Table 17: **RL performance of the 4B Dense Model with $\beta = 0.50$; $\tilde{\lambda} = 0$; $\hat{\lambda} = 0$.**

	# RL Steps	100	200	300	400	500	600	700	800	900	1000
AIME24	Avg@128	0.94	1.38	1.43	1.46	2.29	1.95	1.74	2.19	2.53	2.42
	Cons@128	3.33	3.33	3.33	3.33	6.67	3.33	3.33	3.33	3.33	3.33
	Pass@64	18.77	23.78	16.58	20.87	23.65	19.99	18.98	24.56	23.62	22.08
AIME25	Avg@128	0.83	1.88	1.95	2.03	3.39	3.57	2.76	3.67	4.14	4.66
	Cons@128	0.00	6.67	3.33	3.33	10.00	10.00	3.33	6.67	10.00	10.00
	Pass@64	17.05	15.84	17.51	13.33	15.00	19.59	15.84	20.01	18.77	15.00
AMC23	Avg@128	19.34	22.52	23.09	26.35	26.60	25.46	28.28	29.45	29.49	30.10
	Cons@128	32.50	37.50	37.50	40.00	45.00	42.50	37.50	40.00	42.50	42.50
	Pass@64	76.20	73.68	75.33	71.56	72.32	74.94	68.79	71.99	73.76	73.35
OlympiadBench	Avg@128	14.37	18.07	20.35	21.00	22.38	22.27	22.24	23.32	24.44	24.96
	Cons@128	23.56	26.22	28.15	28.59	30.37	29.93	29.33	30.81	31.85	32.00
	Pass@64	54.84	57.02	57.95	56.39	57.25	58.01	55.62	57.67	57.93	56.91
MATH-500	Avg@128	39.40	45.70	48.18	49.46	50.41	50.58	51.53	52.32	53.20	54.00
	Cons@128	56.00	59.40	61.80	60.20	61.40	61.80	60.20	62.20	62.40	63.20
	Pass@64	85.72	87.81	88.00	88.72	88.92	89.49	88.58	88.58	89.39	89.76
Minerva	Avg@128	8.50	10.44	11.41	11.34	11.96	11.67	11.73	11.71	12.23	12.10
	Cons@128	16.91	16.91	17.65	18.01	18.01	16.91	16.54	15.44	15.44	16.18
	Pass@64	42.52	45.32	44.50	43.82	44.10	43.17	41.51	45.40	46.06	43.45
Average	Avg@128	13.90	16.67	17.74	18.61	19.51	19.25	19.71	20.44	21.01	21.37
	Cons@128	22.05	25.01	25.29	25.58	28.58	27.41	25.04	26.41	27.59	27.87
	Pass@64	49.18	50.58	49.98	49.12	50.21	50.87	48.22	51.37	51.59	50.09

Table 18: **RL performance of the 10B-A0.5B MoE Model with** $\beta = -0.25;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$**.**

	# RL Steps	100	200	300	400	500	600	700	800	900	1000
AIME24	Avg@128	0.29	0.49	0.86	0.49	0.49	0.86	0.86	0.78	0.86	1.15
	Cons@128	0.00	0.00	3.33	0.00	0.00	3.33	0.00	3.33	0.00	3.33
	Pass@64	9.15	7.93	8.13	10.84	11.68	14.17	14.59	19.19	15.01	21.27
AIME25	Avg@128	0.16	0.29	0.36	0.36	0.63	0.63	0.89	0.70	0.81	1.04
	Cons@128	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Pass@64	8.35	13.78	15.46	7.92	19.81	17.11	19.20	13.12	14.17	17.21
AMC23	Avg@128	11.52	14.02	16.88	18.48	19.47	21.05	22.44	22.66	21.25	22.40
	Cons@128	30.00	35.00	37.50	37.50	37.50	32.50	35.00	35.00	32.50	30.00
	Pass@64	75.75	73.72	73.54	76.83	74.15	70.35	71.18	74.50	74.13	78.66
OlympiadBench	Avg@128	7.08	9.60	11.51	12.79	14.47	15.71	16.85	16.02	16.78	17.68
	Cons@128	13.33	17.78	20.59	20.89	21.93	24.59	24.15	24.00	25.19	26.96
	Pass@64	47.16	49.12	52.08	51.30	50.62	52.86	53.15	52.78	54.66	55.99
MATH-500	Avg@128	24.43	30.21	34.68	37.65	40.62	42.61	44.71	45.35	44.40	44.86
	Cons@128	45.60	48.60	52.60	54.20	54.20	56.20	57.80	59.40	57.00	58.00
	Pass@64	81.19	83.88	85.36	84.99	85.75	86.98	88.10	87.39	87.16	86.89
Minerva	Avg@128	5.26	6.36	7.45	8.24	9.02	9.89	10.29	10.48	10.14	9.96
	Cons@128	10.29	13.24	13.24	13.24	15.07	15.07	15.81	15.81	16.18	16.91
	Pass@64	38.70	41.40	43.51	44.46	43.11	45.72	45.94	45.16	45.82	44.46
Average	Avg@128	8.12	10.16	11.96	13.00	14.12	15.13	16.01	16.00	15.71	16.18
	Cons@128	16.54	19.10	21.21	20.97	21.45	21.95	22.13	22.92	21.81	22.53
	Pass@64	43.38	44.97	46.35	46.06	47.52	47.87	48.69	48.69	48.49	50.75

Table 19: **RL performance of the 10B-A0.5B MoE Model with** $\beta = 0.50;\ \tilde{\lambda} = 0;\ \hat{\lambda} = 0$**.**

	# RL Steps	100	200	300	400	500	600	700	800	900	1000
AIME24	Avg@128	0.42	0.29	0.65	0.70	0.68	0.78	0.68	1.02	0.94	1.04
	Cons@128	0.00	0.00	3.33	3.33	0.00	3.33	0.00	3.33	3.33	3.33
	Pass@64	18.31	8.32	16.69	19.19	18.77	14.17	12.51	17.53	20.66	16.58
AIME25	Avg@128	0.26	0.16	0.18	0.34	0.26	0.36	0.29	0.26	0.34	0.68
	Cons@128	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Pass@64	14.19	9.17	10.01	14.41	13.77	14.97	12.11	11.06	13.67	19.02
AMC23	Avg@128	11.17	13.46	13.54	14.14	15.90	17.27	16.15	17.60	18.42	19.67
	Cons@128	17.50	25.00	22.50	25.00	25.00	30.00	27.50	30.00	32.50	32.50
	Pass@64	70.56	72.46	71.53	78.57	76.84	78.90	79.26	74.73	78.60	77.41
OlympiadBench	Avg@128	7.14	8.17	9.46	10.45	11.60	12.76	12.15	13.38	14.04	15.17
	Cons@128	14.81	15.56	16.59	19.11	20.44	21.63	20.44	21.78	22.81	25.19
	Pass@64	47.23	50.58	51.54	51.79	52.53	51.78	52.20	52.60	53.27	52.94
MATH-500	Avg@128	27.15	30.51	33.40	34.80	36.67	38.32	36.86	39.42	40.66	41.36
	Cons@128	44.40	47.40	47.80	50.60	52.00	52.80	52.20	55.00	55.60	57.00
	Pass@64	81.18	82.20	82.85	83.56	84.79	84.78	85.65	86.00	85.97	86.24
Minerva	Avg@128	5.86	5.92	6.83	7.01	7.51	8.13	7.94	8.38	8.80	9.01
	Cons@128	11.40	10.66	11.76	11.03	12.13	12.13	12.50	13.97	13.60	16.91
	Pass@64	41.33	42.59	43.28	42.82	44.07	44.88	44.56	44.13	46.17	45.37
Average	Avg@128	8.67	9.75	10.68	11.24	12.10	12.94	12.35	13.34	13.87	14.49
	Cons@128	14.69	16.44	17.00	18.18	18.26	19.98	18.77	20.68	21.31	22.49
	Pass@64	45.47	44.22	45.98	48.39	48.46	48.25	47.72	47.68	49.72	49.59

Table 20: **RL performance of the 4B Dense Model with $\beta = 0$; $\tilde{\lambda} = 0$; $\hat{\lambda} = -0.1$.**

	# RL Steps	100	200	300	400	500	600	700	800	900	1000
AIME24	Avg@128	0.70	0.70	0.76	0.89	1.07	0.81	0.86	1.15	1.38	1.46
	Cons@128	0.00	3.33	3.33	3.33	3.33	3.33	3.33	3.33	3.33	3.33
	Pass@64	22.51	12.51	15.01	14.19	13.77	15.01	20.87	21.03	22.93	23.70
AIME25	Avg@128	0.76	1.09	1.46	1.90	1.67	2.32	2.11	3.02	3.72	3.18
	Cons@128	0.00	0.00	0.00	3.33	3.33	6.67	3.33	3.33	6.67	3.33
	Pass@64	13.97	15.85	17.50	19.59	18.29	16.68	15.00	21.90	23.58	18.33
AMC23	Avg@128	18.50	20.76	21.84	24.67	21.82	25.35	27.55	28.73	30.47	31.13
	Cons@128	37.50	40.00	37.50	40.00	40.00	37.50	37.50	42.50	42.50	45.00
	Pass@64	76.21	78.48	79.17	73.11	69.87	77.31	72.07	75.05	73.18	77.07
OlympiadBench	Avg@128	13.79	15.29	17.63	19.50	18.03	19.51	20.84	22.23	23.58	24.72
	Cons@128	22.22	24.30	25.04	26.81	27.11	27.41	27.41	29.63	29.78	31.70
	Pass@64	54.56	56.07	55.14	55.54	56.67	56.07	55.65	56.93	57.90	56.66
MATH-500	Avg@128	37.92	40.85	43.98	45.69	43.78	46.94	48.87	50.44	51.11	53.30
	Cons@128	56.20	58.20	59.00	59.00	59.80	59.60	61.80	63.00	64.00	65.00
	Pass@64	86.56	87.15	87.23	87.95	87.60	87.54	87.25	88.08	89.22	89.02
Minerva	Avg@128	8.07	9.22	10.25	11.08	10.67	11.43	12.28	12.54	12.24	13.13
	Cons@128	17.28	18.01	15.07	15.81	17.28	16.18	16.54	18.01	18.38	18.38
	Pass@64	41.61	45.66	46.22	43.84	43.71	44.43	44.29	43.08	41.96	42.28
Average	Avg@128	13.29	14.65	15.99	17.29	16.17	17.73	18.75	19.69	20.42	21.25
	Cons@128	22.20	23.97	23.32	24.71	25.14	25.12	24.99	26.63	27.44	27.79
	Pass@64	49.24	49.29	50.05	49.04	48.32	49.51	49.19	51.01	51.46	51.18

Table 21: **RL performance of the 4B Dense Model with $\beta = 0$; $\tilde{\lambda} = 0.1$; $\hat{\lambda} = 0$.**

	# RL Steps	100	200	300	400	500	600	700	800	900	1000
AIME24	Avg@128	0.49	0.55	0.91	0.73	0.68	0.63	0.68	0.76	0.91	1.28
	Cons@128	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.33
	Pass@64	9.17	13.24	23.37	8.13	18.97	17.10	20.65	17.94	16.37	20.32
AIME25	Avg@128	0.68	0.83	1.35	1.46	1.28	1.25	1.51	1.93	2.29	2.29
	Cons@128	0.00	0.00	3.33	3.33	3.33	3.33	3.33	3.33	3.33	6.67
	Pass@64	21.79	22.09	17.83	20.01	16.61	15.81	17.48	15.84	16.26	15.00
AMC23	Avg@128	18.24	20.59	24.16	25.16	23.89	28.38	29.20	29.57	31.43	32.68
	Cons@128	32.50	32.50	42.50	35.00	42.50	42.50	42.50	37.50	45.00	45.00
	Pass@64	75.30	71.98	75.55	77.70	70.73	71.66	72.38	72.02	75.16	74.30
OlympiadBench	Avg@128	13.38	16.45	18.73	19.53	19.28	21.67	21.92	22.44	23.38	24.35
	Cons@128	22.96	25.48	27.56	27.70	27.11	29.48	29.33	29.93	31.41	32.00
	Pass@64	54.04	55.12	55.80	56.96	56.31	55.42	57.13	56.75	57.94	57.64
MATH-500	Avg@128	37.71	43.26	47.34	48.65	48.09	51.15	51.84	52.49	53.79	54.85
	Cons@128	56.00	58.80	62.20	63.20	63.60	63.80	62.80	64.80	64.60	65.40
	Pass@64	84.83	87.11	88.42	88.69	87.97	87.80	87.65	88.55	88.98	88.76
Minerva	Avg@128	8.00	9.30	10.37	10.83	11.03	11.39	11.47	11.51	11.70	11.66
	Cons@128	14.71	16.54	16.91	18.38	16.91	16.91	16.54	18.01	16.54	15.81
	Pass@64	41.49	43.44	45.78	44.18	44.17	43.23	44.97	44.22	44.80	45.56
Average	Avg@128	13.08	15.16	17.14	17.73	17.38	19.08	19.44	19.78	20.58	21.19
	Cons@128	21.03	22.22	25.42	24.60	25.58	26.00	25.75	25.60	26.81	28.04
	Pass@64	47.77	48.83	51.13	49.28	49.13	48.50	50.04	49.22	49.92	50.26

Table 22: **RL performance of the 10B-A0.5B MoE Model with $\beta = 0;\ \tilde{\lambda} = 0;\ \hat{\lambda} = -0.1$.**

	# RL Steps	100	200	300	400	500	600	700	800	900	1000
AIME24	Avg@128	0.37	0.34	0.50	0.70	0.78	0.70	0.76	0.81	0.78	1.22
	Cons@128	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Pass@64	11.45	11.66	13.77	18.58	26.30	17.94	19.83	23.30	26.39	25.86
AIME25	Avg@128	0.39	0.18	0.31	0.31	0.60	0.52	0.55	0.60	0.55	0.76
	Cons@128	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Pass@64	12.50	10.84	12.83	13.25	19.91	15.01	23.10	15.03	14.19	13.24
AMC23	Avg@128	11.91	14.30	15.12	17.87	19.71	20.27	21.72	22.68	22.87	24.32
	Cons@128	25.00	35.00	27.50	35.00	32.50	37.50	35.00	37.50	35.00	40.00
	Pass@64	72.28	75.78	74.15	75.68	77.69	74.47	75.45	70.93	80.24	73.19
OlympiadBench	Avg@128	8.08	9.71	10.69	12.87	14.36	14.86	15.96	16.31	16.76	17.47
	Cons@128	16.00	20.00	20.89	23.70	23.26	23.26	24.74	25.33	24.59	25.93
	Pass@64	49.35	50.50	52.65	54.15	53.54	53.70	54.05	55.85	54.78	54.95
MATH-500	Avg@128	27.41	32.00	34.17	38.08	40.73	41.59	43.06	42.97	44.00	45.20
	Cons@128	47.20	51.20	53.60	55.00	57.40	58.00	57.80	58.40	59.40	59.00
	Pass@64	84.11	84.77	84.79	86.88	86.96	87.37	86.11	87.32	87.34	87.70
Minerva	Avg@128	5.54	6.92	7.44	8.61	9.29	9.17	9.88	9.82	9.92	10.03
	Cons@128	11.76	15.44	15.81	16.91	15.44	15.81	16.54	16.91	15.81	18.01
	Pass@64	40.01	40.64	41.59	43.33	42.83	44.31	44.58	44.51	44.57	44.25
Average	Avg@128	8.95	10.58	11.37	13.07	14.25	14.52	15.32	15.53	15.81	16.50
	Cons@128	16.66	20.27	19.63	21.77	21.43	22.43	22.35	23.02	22.47	23.82
	Pass@64	44.95	45.70	46.63	48.65	51.21	48.80	50.52	49.49	51.25	49.87

Table 23: **RL performance of the 10B-A0.5B MoE Model with $\beta = 0;\ \tilde{\lambda} = 0.1;\ \hat{\lambda} = 0$.**

	# RL Steps	100	200	300	400	500	600	700	800	900	1000
AIME24	Avg@128	0.32	0.50	0.65	0.86	0.83	0.96	0.81	0.68	0.83	0.99
	Cons@128	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Pass@64	9.57	10.84	9.17	21.66	10.84	13.96	16.48	10.85	13.67	20.77
AIME25	Avg@128	0.21	0.26	0.34	0.44	0.42	0.76	0.83	0.99	1.17	1.02
	Cons@128	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.33	3.33
	Pass@64	8.97	10.65	14.62	16.81	11.14	18.23	18.34	16.27	17.94	17.51
AMC23	Avg@128	12.30	14.57	17.58	19.16	19.90	20.88	20.55	21.35	20.63	19.86
	Cons@128	27.50	30.00	30.00	32.50	30.00	37.5	35.00	40.00	35.00	32.50
	Pass@64	71.35	67.79	73.48	73.13	74.08	75.33	77.74	71.64	78.56	76.32
OlympiadBench	Avg@128	8.81	11.18	13.36	14.83	15.89	16.77	17.09	16.27	16.30	16.13
	Cons@128	15.70	19.11	21.78	23.85	24.59	25.04	24.89	24.44	24.00	24.30
	Pass@64	49.50	49.62	50.93	50.98	52.84	52.59	53.86	51.74	51.80	52.61
MATH-500	Avg@128	28.13	33.38	37.19	39.12	41.51	42.88	43.12	43.56	43.53	42.68
	Cons@128	48.00	50.40	52.00	53.20	57.00	57.40	57.20	58.40	58.40	58.80
	Pass@64	83.25	83.85	86.04	85.39	85.14	85.43	85.67	85.36	86.33	86.73
Minerva	Avg@128	5.50	6.58	7.76	8.31	9.08	8.88	8.95	7.83	7.89	7.78
	Cons@128	8.46	12.13	11.40	12.50	15.44	15.07	13.97	11.03	9.93	11.40
	Pass@64	40.19	40.53	42.63	43.57	44.28	43.39	44.42	42.66	42.90	42.55
Average	Avg@128	9.21	11.08	12.81	13.79	14.61	15.19	15.23	15.11	15.06	14.74
	Cons@128	16.61	18.61	19.20	20.34	21.17	22.50	21.84	22.31	21.78	21.72
	Pass@64	43.81	43.88	46.15	48.59	46.39	48.16	49.42	46.42	48.53	49.42