

Bottom-up Policy Optimization: Your Language Model Policy Secretly Contains Internal Policies

Yuqiao Tan^{1,2*}, Minzheng Wang^{1,2*}, Shizhu He^{1,2†},
Huanxuan Liao^{1,2}, Chengfeng Zhao^{1,2}, Liang Tian³, Jun Zhao^{1,2}, Kang Liu^{1,2}

¹ The Key Laboratory of Cognition and Decision Intelligence for Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China

² School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

³ Tencent AI Lab

{tanyuqiao2025, wangminzheng2023}@ia.ac.cn {shizhu.he, jzhao, kliu}@nlpr.ia.ac.cn

Abstract

Existing reinforcement learning (RL) approaches treat large language models (LLMs) as a unified policy, overlooking their internal mechanisms. In this paper, we decompose the LLM-based policy into **Internal Layer Policies** and **Internal Modular Policies** via the Transformer’s residual stream. Our entropy analysis of internal policy reveals distinct patterns: (1) universally, internal policies evolve from high-entropy exploration in early layers to deterministic refinement in the top layers; and (2) Qwen exhibits an explicit progressive reasoning structure, contrasting with the abrupt convergence in Llama. Furthermore, we discover that optimizing internal layers induces feature refinement, forcing lower layers to capture high-level reasoning representations early. Motivated by these findings, we propose **Bottom-up Policy Optimization (BuPO)**, a novel RL paradigm that reconstructs the LLM’s reasoning foundation from the bottom up by optimizing internal layers in early stages. Extensive experiments on complex reasoning benchmarks demonstrate the effectiveness of BuPO. Our code is available at <https://github.com/Trae1ounG>.

1 Introduction

Reinforcement learning (RL) has become central to improving the complex reasoning abilities of large language models (LLMs) (Ouyang et al., 2022; Jaech et al., 2024). The success of DeepSeek-R1 (Guo et al., 2025) further highlights reinforcement learning with verifiable rewards (RLVR) as an effective post-training paradigm for improving LLM reasoning across domains (Yang et al., 2025; Team et al., 2025; Yu et al., 2025). However, existing RL methods typically treat an LLM as a unified policy, optimizing only the final output distribution while overlooking the evolving information flow

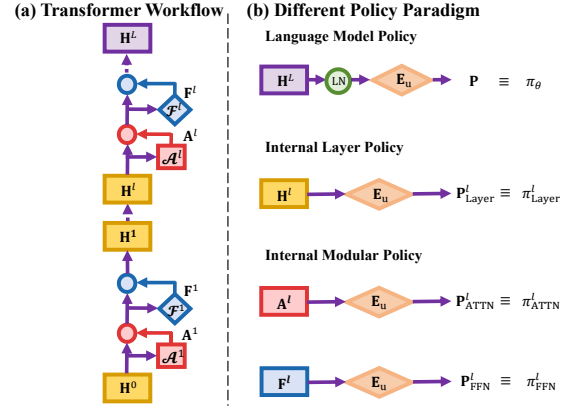


Figure 1: (a) The residual stream within a Transformer flows from previous layer into self-attention and feed-forward network (FFN) sequentially. (b) Any hidden states combined with the unembedding matrix E_u can be transformed into probability distribution P over the vocabulary space, which can be considered a policy.

in its internal residual streams. Mechanistic interpretability provides tools for exposing such internal behavior (Belrose et al., 2023; Gupta et al., 2025), creating an opportunity to connect internal dynamics with more principled policy optimization.

While recent works leverage attention mechanisms to improve RL algorithms (Li et al., 2025; Liu et al., 2025), they largely overlook the information latent in residual streams. Logit-lens (nostalgebraist, 2020) offers initial insights by using the unembedding matrix E_u to decode intermediate layer representations into token space, revealing rich information that evolves across layers and modules (Gupta et al., 2025; Lindsey et al., 2025). Moreover, prior studies have further elucidated the roles of self-attention and feed-forward networks (FFNs) in shaping hidden states (Dai et al., 2022; Yu and Ananiadou, 2024; Jin et al., 2025). Collectively, these internal mechanisms offer a new perspective for algorithmic optimization.

In this paper, we investigate the evolution of language model policies across layers and modules

*Equal contribution.

†Corresponding author

to facilitate optimization and unravel complex internal reasoning mechanisms. Our formulation is grounded in two key insights. First, the residual stream naturally supports decomposition (Zhang et al., 2025; Lindsey et al., 2025), allowing us to isolate the individual roles of each layer and module (Figure 1(a)). Second, we conclude that the policy is intrinsically equivalent to the token distribution derived from the combination of hidden states $\mathbf{H}$ with the unembedding matrix $\mathbf{E}_u$. Based on these, we construct the **Internal Layer Policy** π_{Layer}^l , which captures cumulative reasoning up to layer l , and the **Internal Modular Policy** π_{ATTN}^l and π_{FFN}^l , which isolates the specific contributions of attention and FFN modules (Figure 1(b)). This decomposition allows us to ask: *How does internal reasoning evolve through the model?*

Through systematic analysis of commonly used Qwen and Llama series (Meta AI, 2024; Yang et al., 2024, 2025) based on **Internal Policy Entropy** in a policy-centric view, we uncover both universal and critical architectural differences: (1) **Consistent internal reasoning structure**. All models exhibit a universal reasoning structure: early layers maintain high entropy for exploring the solution space, while top layers converge to near-zero for final prediction (Lindsey et al., 2025). (2) **Distinct internal reasoning pattern**. Despite the shared trend, the pace of convergence differs significantly. Llama exhibits a sudden convergence only within the last three layers. In contrast, Qwen models demonstrate a progressive contraction, gradually reducing uncertainty throughout layers. To quantify these dynamics, we introduce **Entropy Change**. This metric shows that Llama exhibits increased internal entropy and tends to converge only in the final layer, whereas Qwen progressively leverages FFNs to expand exploration in lower layers, integrate parametric knowledge in intermediate layers, and consolidate predictions in upper layers.

These findings have profound implications for RL optimization: *Since internal reasoning emerges from lower to higher, we can consider optimization from a bottom-up perspective*. We first validate this hypothesis with Internal Policy Alignment (IPA), revealing distinct training dynamics and a remarkable phenomenon of internal reasoning feature refinement. Specifically, the optimized lower layers capture high-level reasoning capabilities by early alignment, providing a robust foundation for subsequent internal reasoning. Motivated by these

insights, we propose **Bottom-up Policy Optimization (BuPO)**, a novel RL paradigm that optimizes fine-grained internal layer policies during the early stages of training to effectively guide the language model policy. By doing so, BuPO reconstructs foundational reasoning abilities and achieves superior performance. Extensive experiments on complex reasoning benchmarks demonstrate the effectiveness of our approach and the unique training dynamics compared to conventional RL methods that optimize the policy as a whole.

In summary, this paper makes the following contributions: (1) We are the first to decompose an LLM policy into internal layer and modular policies, revealing their distinct roles in reasoning. (2) We reveal a universal exploration-to-convergence shift, distinguishing Qwen’s progressive reasoning from Llama’s abrupt convergence. Moreover, we uncover a feature refinement phenomenon where internal optimization drives lower layers to preemptively capture high-level features. (3) We propose BuPO to align internal policies from the bottom up, which reconstructs the foundational reasoning capabilities at lower layers, achieving superior performance on complex reasoning benchmarks.

2 Preliminary

In this section, we aim to introduce the basic definitions that help us to understand the decomposition of the language model policy.

2.1 The Residual Stream in Transformer

Transformer-based language models (Vaswani et al., 2017) form the foundation of modern LLMs (Brown et al., 2020). A decoder-only Transformer consists of L stacked layers, each containing a multi-head self-attention (MHSA) module and a feed-forward network (FFN) module.

Following Zhang et al. (2025), we formalize the forward process from input to output. Given an input sequence $\mathbf{x} = [x_1, x_2, \dots, x_T]$, the model produces a probability distribution $\mathbf{P}$ over the vocabulary V with N tokens. Let $\mathbf{H}^{(2l-2)} \in \mathbb{R}^{T \times d_{\text{model}}}$ denote the hidden state input to the l -th layer, where T is the sequence length and d_{model} is the hidden dimension. The initial state is $\mathbf{H}^{(0)}$ projected by $\mathbf{E}$, where $\mathbf{E} \in \mathbb{R}^{N \times d_{\text{model}}}$ is the embedding matrix.

Each layer forwards sequentially through attention and FFN with residual connections: $\mathbf{H}^{(2l-1)} = \mathbf{H}^{(2l-2)} + \mathbf{A}^l$ and $\mathbf{H}^{(2l)} = \mathbf{H}^{(2l-1)} + \mathbf{F}^l$, where $\mathbf{A}^l = \text{MHSA}(\text{LN}(\mathbf{H}^{(2l-2)}))$ and $\mathbf{F}^l =$

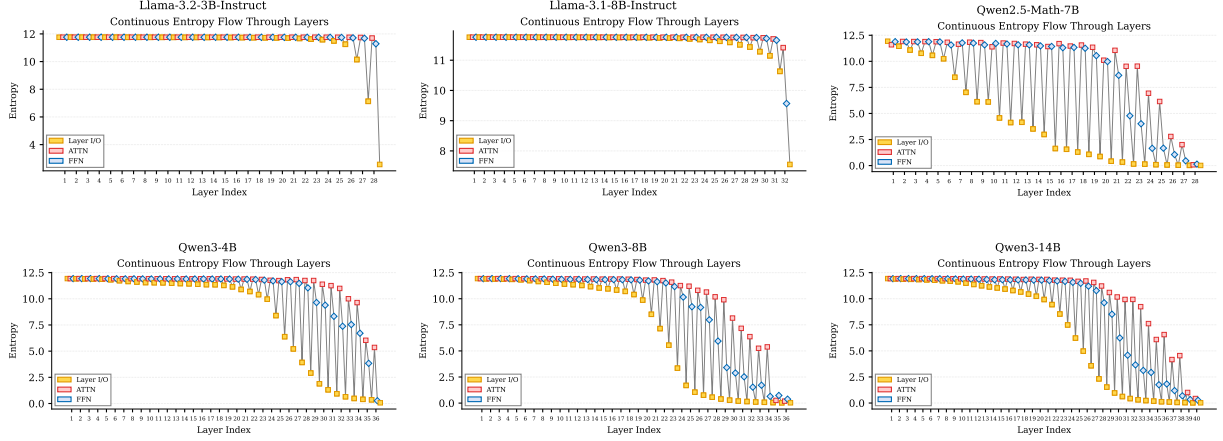


Figure 2: Continuous entropy dynamics of internal policy. The residual stream flows from $\mathbf{H}^{l-1}$ into $\mathbf{A}^l$, $\mathbf{F}^l$, then to the next layer $\mathbf{H}^l$.

$\text{FFN}(\text{LN}(\mathbf{H}^{(2l-1)}))$ are the attention and FFN outputs, and $\text{LN}(\cdot)$ denotes layer normalization.

After L layers, the final hidden states are projected to vocabulary logits:

$$\mathbf{P} = \text{softmax}(\text{LN}(\mathbf{H}^{(2L)})\mathbf{E}_u^T), \quad (1)$$

where $\mathbf{E}_u \in \mathbb{R}^{N \times d_{\text{model}}}$ is the unembedding matrix, and $\mathbf{P} \in \mathbb{R}^{T \times N}$ denotes the output distribution.

2.2 Reinforcement Learning for Language Model Policy

Language model generation can be formulated as a token-level Markov Decision Process (MDP). At each step t , the state $s_t = [\mathbf{q}; \mathbf{o}_{<t}]$ consists of the input and generated tokens so far. The language model policy $\pi_\theta(\cdot|s_t)$ samples the next token o_t from vocabulary V , transitioning to $s_{t+1} = [s_t; o_t]$. To optimize the policy, we maximize:

$$\mathcal{J}(\pi_\theta) = \mathbb{E}_{\mathbf{q} \sim \mathcal{Q}, \mathbf{o} \sim \pi_\theta(\cdot|\mathbf{q})} [R(\mathbf{q}, \mathbf{o}) - \beta \mathbb{D}_{KL}[\pi_\theta(\cdot|\mathbf{q}) || \pi_{\text{ref}}(\cdot|\mathbf{q})]], \quad (2)$$

where $R(\mathbf{q}, \mathbf{o}) = \sum_{t=1}^{|\mathbf{o}|} r(s_t, o_t)$ is the return (Sutton et al., 1998) and π_{ref} is a reference policy. We adopt sparse rewards where $r_t = 0$ for $t < n$ and $r_n \in [0, 1]$ indicates task success. Following Hu et al. (2025b), we assume $\beta = 0$.

Policy Optimization. We adopt GRPO (Shao et al., 2024), which samples a group of responses $\{\mathbf{o}_1, \dots, \mathbf{o}_G\}$ per question and estimates advan-

tages as $\hat{A}_{i,t} = \frac{R_i - \text{mean}(\mathbf{R})}{\text{std}(\mathbf{R})}$:

$$\mathcal{J}_{\text{GRPO}}(\pi_\theta) = \mathbb{E}_{\mathbf{q} \sim \mathcal{Q}, \{\mathbf{o}_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|\mathbf{q})} \frac{1}{G} \sum_{i=1}^G \frac{1}{|\mathbf{o}_i|} \sum_{t=1}^{|\mathbf{o}_i|} \left\{ \min \left[r_{i,t} \hat{A}_{i,t}, \text{clip}(r_{i,t}, 1 - \epsilon, 1 + \epsilon) \hat{A}_{i,t} \right] \right\}, \quad (3)$$

where $r_{i,t} = \frac{\pi_\theta(o_{i,t}|s_{i,t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|s_{i,t})}$ is the importance ratio.

3 Language Model Policy Secretly Contains Internal Policies

In this section, we introduce our key insight: *the language model policy secretly contains internal policies*. We provide implementation details of this section in Appendix A.3.

3.1 Definition of Internal Policy

Residual Stream. In the residual stream of Transformer, the input to any layer equals the sum of all preceding outputs plus the original embedding. The hidden states satisfy:

$$\mathbf{H}^l = \mathbf{H}^{(2l)} = \mathbf{H}^{(0)} + \sum_{i=1}^l \mathbf{A}^i + \sum_{j=1}^l \mathbf{F}^j, \quad (4)$$

where we denote $\mathbf{H}^l$ for simplicity as the output hidden states of layer l , $\mathbf{H}^{(2l)}$. According to this, the output of final layer can be regarded as the combination of previous hidden states by $\mathbf{H}^L = \mathbf{H}^0 + \sum_{i=1}^L \mathbf{A}^i + \sum_{j=1}^L \mathbf{F}^j$.

Internal Policy. During RL, we sample the next token o_t from the final layer’s probability distribution, i.e., $\pi_\theta \equiv \mathbf{P} = \text{softmax}(\text{LN}(\mathbf{H}^L)\mathbf{E}_u^T)$. We

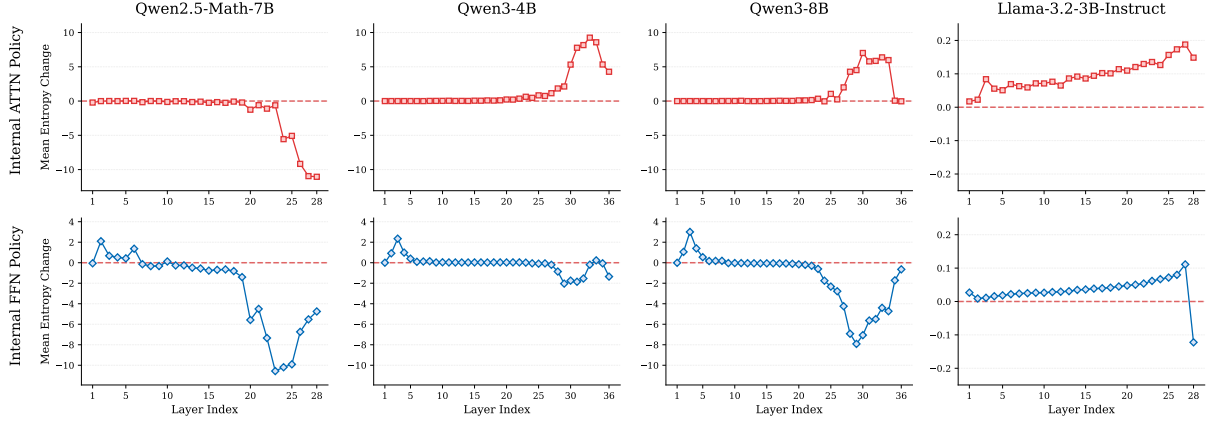


Figure 3: Entropy change dynamics of internal policy. The entropy change ΔH^l across layers represents the uncertainty of current policy’s hidden exploration space. $\Delta H^l > 0$ indicates increased exploration, $\Delta H^l \approx 0$ signifies exploitation of existing knowledge, and $\Delta H^l < 0$ suggests convergence within the reasoning process.

propose that each internal hidden states can be combined with $\mathbf{E}_u$ to produce a policy. Specifically, we focus on two granularities. **Internal Layer Policy** refers to using hidden states from each layer $\mathbf{H}^l$ to combine with $\mathbf{E}_u$, and **Internal Modular Policy** integrates $\mathbf{E}_u$ with states from specific module:

$$\pi_{\text{Layer}}^l \equiv \mathbf{P}_{\text{Layer}}^l = \text{softmax}(\mathbf{H}^l \mathbf{E}_u^T), \quad (5)$$

$$\pi_{\text{Module}}^l \equiv \begin{cases} \mathbf{P}_{\text{ATTN}}^l = \text{softmax}(\mathbf{A}^l \mathbf{E}_u^T), & \text{for ATTN} \\ \mathbf{P}_{\text{FFN}}^l = \text{softmax}(\mathbf{F}^l \mathbf{E}_u^T). & \text{for FFN} \end{cases} \quad (6)$$

Each component contributes to the final policy through the residual stream. For instance, $\mathbf{H}^L = \mathbf{H}^l + \mathbf{S}^{l+1}$, where $\mathbf{S}^{l+1} = \sum_{i=l+1}^L \mathbf{A}^i + \sum_{j=l+1}^L \mathbf{F}^j$ represents contributions from subsequent layers. Hence, understanding these internal components is essential for unraveling *how internal reasoning emerges and evolves through the model*.

3.2 Internal Policy Entropy Dynamics

In contrast to prior logit-lens approaches (nostalgabraist, 2020) that decode internal states into discrete tokens, we adopt a policy-centric perspective where internal probability distributions are treated as policies. We employ **Entropy** as our primary probing metric, motivated by its strong correlation with policy behavior (Cui et al., 2025; Cheng et al., 2025). We define **Internal Policy Entropy** as:

$$H_{\text{Layer}}^l = - \sum_{j=1}^{|V|} \mathbf{P}_{\text{Layer},j}^l \cdot \log(\mathbf{P}_{\text{Layer},j}^l), \quad (7)$$

where $|V|$ denotes the vocabulary size and we can obtain H_{FFN}^l and H_{ATTN}^l in the same way.

Continuous Entropy Dynamics. Figure 2 shows that internal policy entropy dynamics exhibit consistent patterns across models: early layers maintain high entropy for exploration of the search space, while top layers converge to near-zero entropy. This aligns with findings that lower layers capture semantic information while higher layers aggregate and refine these representations to drive final decision-making (Lindsey et al., 2025).

While the overall entropy pattern is consistent across models, the fine-grained transition dynamics vary. To isolate intrinsic patterns from normalization and residual effects (He et al., 2016; Zhang and Sennrich, 2019), we introduce **Entropy Change**, which measures the incremental information gain within a single internal policy and is defined as:

$$\Delta H^l = H_{\text{Output}}^l - H_{\text{Input}}^l, \quad (8)$$

where the entropy change is defined as the difference between the internal policy entropy at a module’s input and output. This metric reveals how the exploration space evolves as information propagates through the specific module.

Entropy Change Dynamics of Attention vs. FFN. *Self-attention* modules (Vaswani et al., 2017) are widely regarded as central to model reasoning, particularly for integrating task-relevant contextual information (Jin et al., 2025; Liu et al., 2025). The upper panel of Figure 3 reveals a clear and model-dependent pattern in the entropy change of self-attention. Specifically, Qwen3 models exhibit consistently positive entropy change across layers ($\Delta H_{\text{ATTN}}^l > 0$), indicating sustained expansion of the exploration space during reasoning. In contrast,

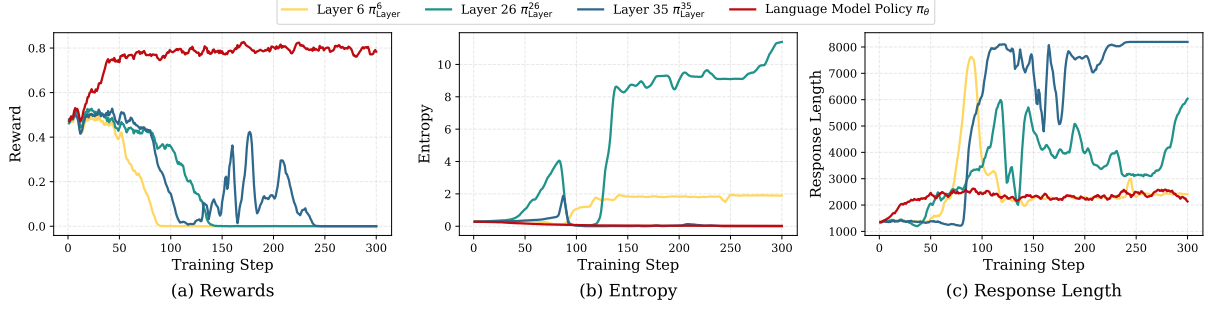


Figure 4: Training dynamics of internal policy. Effects of varying the optimized policy on (a) reward, (b) entropy of language model policy π_θ , (c) response length. The backbone model is Qwen3-4B.

Qwen2.5-Math-7B shows uniformly negative entropy change, suggesting contraction and earlier convergence driven by attention. Llama models display a weaker but still positive trend, reflecting more conservative exploration dynamics.

Moreover, the *FFN* modules are widely regarded as the key-value memories of parametric knowledge (Geva et al., 2021; Dai et al., 2022). The lower panel of Figure 3 reveals clear and systematic differences in *FFN* entropy dynamics across model families. For the Llama models (Meta AI, 2024), *FFN* entropy remains consistently positive but weak across almost all layers, with convergence occurring only at the final layer.

By contrast, the Qwen models (Yang et al., 2024, 2025) exhibit a pronounced hierarchical entropy structure in the *FFN*, following a clear three-stage progression. Taking Qwen3-4B as an example, the lower *FFN* layers (layers 1–6) show increased entropy ($\Delta H_{\text{FFN}}^l > 0$), corresponding to expanded exploration at the onset of reasoning. This is followed by a broad middle region (layers 7–26) where $\Delta H_{\text{FFN}}^l \approx 0$, indicating stable information integration through retrieval and reuse of parametric knowledge (Dai et al., 2022). In the upper layers (layers 27–36) where $\Delta H_{\text{FFN}}^l < 0$, reflects gradual convergence toward the final prediction.

4 Internal Policy Alignment

Building on the preceding analysis, we observe a gradual emergence of internal reasoning: layer-induced policy distributions exhibit distinct roles and entropy dynamics for different models. This motivates a bottom-up question: *can an internal policy distribution be refined before optimizing the full policy?* We study this through *Internal Policy Alignment* (IPA), an advantage-weighted clipped

surrogate for a selected internal layer policy:

$$\mathcal{J}_{\text{IPA}}(\pi_\theta, \pi_{\text{Layer}}^l) = \mathbb{E}_{\mathbf{q} \sim \mathcal{Q}, \{\mathbf{o}_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|\mathbf{q})} \frac{1}{G} \sum_{i=1}^G \frac{1}{|\mathbf{o}_i|} \sum_{t=1}^{|\mathbf{o}_i|} \min \left[\rho_{i,t}^l \hat{A}_{i,t}, \text{clip}(\rho_{i,t}^l, 1 - \epsilon, 1 + \epsilon) \hat{A}_{i,t} \right], \quad (9)$$

where $\rho_{i,t}^l = \frac{\pi_{\text{Layer}}^l(o_{i,t}|\mathbf{q}, \mathbf{o}_{i,<t})}{\pi_{\text{Layer,old}}(o_{i,t}|\mathbf{q}, \mathbf{o}_{i,<t})}$. The sample process and $\hat{A}_{i,t}$ are the same as in Eq. 3. Thus, $\rho_{i,t}^l$ is not an importance correction for the rollout distribution, but an internal policy update ratio that limits the change of the selected layer-induced distribution on reward-driven tokens. IPA transfers reward feedback to intermediate representations while retaining the stability of clipped policy updates. Implementation details are in Appendix A.4.

Different Training Dynamics of Internal Policy.

As shown in Figure 4, distinct patterns emerge across optimization of different internal layer policies. For the penultimate layer policy π_{Layer}^{35} , entropy shows minor fluctuations before aligning with π_θ . However, it suffers from repetition causing excessively long responses. In contrast, the last integration region policy π_{Layer}^{26} exhibits unstable and increased entropy. And the last exploration region policy π_{Layer}^6 maintains stable entropy growth, with response lengths converging closer to π_θ .

Analysis of Internal Policy Optimization. We further investigate the mechanism behind internal policy alignment. We find that it induces clear feature refinement in internal states. Taking π_{Layer}^6 as an example, Figure 5(a) shows that the optimized $\mathbf{H}^6$ becomes increasingly similar to final layer representations, suggesting that early layers acquire higher-level reasoning features before full policy optimization. Meanwhile, the entropy change in

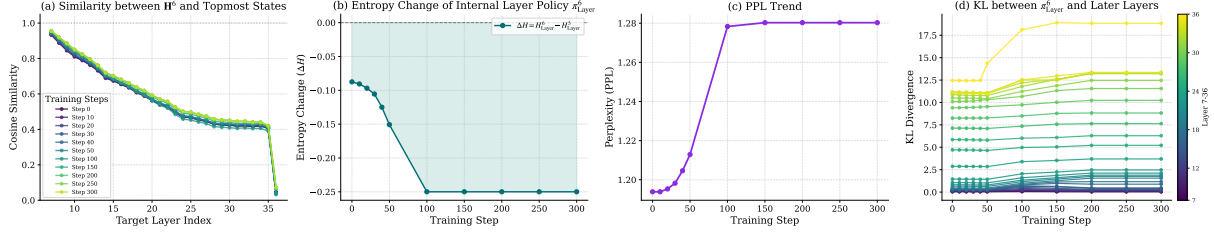


Figure 5: Analysis of internal policy optimization. (a) Similarity between the hidden states of optimized layer 6 and the higher layers. (b) Entropy change $\Delta H^6_{\text{Layer}}$ of the optimized internal layer policy π^6_{Layer} . (c) The PPL trend of π_θ . (d) KL divergence trends between the optimized π^6_{Layer} and later layer policies. The backbone model is Qwen3-4B.

Figure 5(b) shows progressive convergence of the optimized internal policy. Figure 5(c,d) reveals the trade-off of internal policy alignment: while the KL trends show that π^6_{Layer} remains well aligned with later-layer policies during the early stage, this alignment deteriorates with prolonged optimization and coincides with PPL degradation, suggesting that limited internal alignment steps are preferable.

5 Bottom-up Policy Optimization

All prior RL methods for LLMs optimize π_θ in a holistic manner (Ouyang et al., 2022; Rafailov et al., 2023; Guo et al., 2025). In Section 3, we decompose the final output $\mathbf{H}^L$ into an intermediate representation $\mathbf{H}^l$ and the subsequent residual contribution $\mathbf{S}^{l+1}$. This decomposition suggests that aligning the internal policy π^l_{Layer} associated with $\mathbf{H}^l$ may facilitate alignment of the overall policy π_θ . We empirically support this intuition in Section 4, where we optimize the internal policy alone and observe pronounced feature refinement in lower layers. These findings further motivate a bottom-up alignment strategy: by aligning internal policies first, we explore whether the overall policy can be guided to reason more effectively.

To this end, we propose **Bottom-up Policy Optimization (BuPO)**, which sequentially optimizes the internal layer policies π^l_{Layer} followed by the language model policy π_θ :

$$\mathcal{J}_{\text{BuPO}} = \begin{cases} \mathcal{J}_{\text{IPA}}, & s_{\text{cur}} \leq s_{\text{inter}}, \\ \mathcal{J}_{\text{GRPO}}, & s_{\text{cur}} > s_{\text{inter}}, \end{cases} \quad (10)$$

where s_{cur} denotes the current training step, s_{inter} the training steps of the internal layer policy.

Training Setup. Using Qwen (Yang et al., 2025) and Llama (Grattafiori et al., 2024; Wang et al., 2025b) backbones, we assign π^6_{Layer} for both Qwen3-4B and Qwen3-8B, π^{27}_{Layer} for Llama-OctoThinker-3B-Base, and π^{31}_{Layer} for

Llama-OctoThinker-8B-Base. For BuPO, we use the entropy-indicated target layer as the internal policy to optimize, and discuss this choice through the layer-policy ablation in Section 5.3. Refer to Appendix A.5 for detailed setup.

Evaluation Setup. We evaluate BuPO against several RL baselines, including GRPO, PPO (Sutton et al., 1998), Reinforce++ (Hu et al., 2025a) and RLOO (Ahmadian et al., 2024). The benchmarks cover: MATH (Lightman et al., 2024), AMC23 (MAA, 2023), AIME24, and AIME25 (MAA, 2024, 2025). Due to high output variance in reasoning tasks, we report Avg@K (Pass@1 averaged over K outputs). For AIME24/25, we set $K = 32$, and for others $K = 16$. Additionally, we evaluate an unbiased Pass@K metric for a comprehensive evaluation. Detailed evaluation setups are in Appendix A.5.

5.1 Main Results

As shown in Table 1, BuPO consistently improves over baselines across models and benchmarks. On Qwen3-4B, BuPO achieves gains of 4.69 and 2.30 points on AIME24 and AIME25, respectively, with similar improvements on Qwen3-8B (+4.58 and +0.76). The Llama series further validates this trend, with average improvements of 1.01 points on Llama-OctoThinker-3B-Base and 3.68 points on Llama-OctoThinker-8B-Base. We further evaluate BuPO’s generalization on out-of-domain benchmarks spanning general reasoning and coding in Appendix B.1. These results demonstrate that aligning internal layer policies provides a consistent and effective training signal for improving reasoning.

For a comprehensive evaluation, we report the averaged Pass@K results of BuPO and GRPO, with K ranging from 1 to 256. In Figure 6, BuPO consistently achieves a favorable trade-off across a wide range of K . On Qwen3-8B, BuPO attains the best performance for all K values, while on

Methods	AMC (Avg@16)	MATH500 (Avg@16)	AIME24 (Avg@32)	AIME25 (Avg@32)	Average
Qwen3-4B					
Vanilla	67.66	80.29	23.20	18.60	47.44
PPO	77.03	<u>83.64</u>	<u>32.60</u>	27.60	<u>55.22</u>
Reinforce++	63.44	80.63	17.40	18.65	45.03
RLOO	<u>77.66</u>	82.73	30.83	24.79	54.00
GRPO	76.88	82.41	32.19	<u>28.85</u>	55.08
BuPO	81.09+4.21	84.90+2.49	36.88+4.69	31.15+2.30	58.51+3.43
Qwen3-8B					
Vanilla	67.34	80.46	26.98	19.17	48.49
PPO	87.03	86.20	37.81	22.60	58.41
Reinforce++	82.66	86.05	41.77	31.15	60.41
RLOO	86.41	87.32	46.67	33.02	63.36
GRPO	85.94	88.05	49.48	<u>33.54</u>	64.23
BuPO	89.22+3.28	<u>87.76</u>	54.06+4.58	34.38+0.76	66.36+2.13
Llama-OctoThinker-3B-Base					
Vanilla	1.24	5.26	0.21	0.00	1.68
PPO	22.19	43.23	<u>1.04</u>	0.31	16.69
Reinforce++	9.38	11.59	0.00	0.10	5.27
RLOO	<u>27.03</u>	41.93	2.19	0.21	17.84
GRPO	27.50	46.07	0.63	0.10	18.58
BuPO	27.50+0.00	49.79+3.72	0.63+0.00	0.42+0.32	19.59+1.01
Llama-OctoThinker-8B-Base					
Vanilla	4.53	9.84	0.52	0.10	3.75
PPO	31.72	56.97	1.56	1.04	22.82
Reinforce++	34.69	59.55	7.72	<u>3.75</u>	<u>26.43</u>
RLOO	27.66	55.97	3.54	1.56	22.18
GRPO	<u>34.84</u>	56.89	2.50	2.19	24.11
BuPO	37.66+2.82	62.05+5.16	4.69+2.19	6.77+4.58	27.79+3.68

Table 1: Avg@ K results on MATH500, AMC23, AIME24 and AIME25. **Bold** and underlined denote the best and second best.

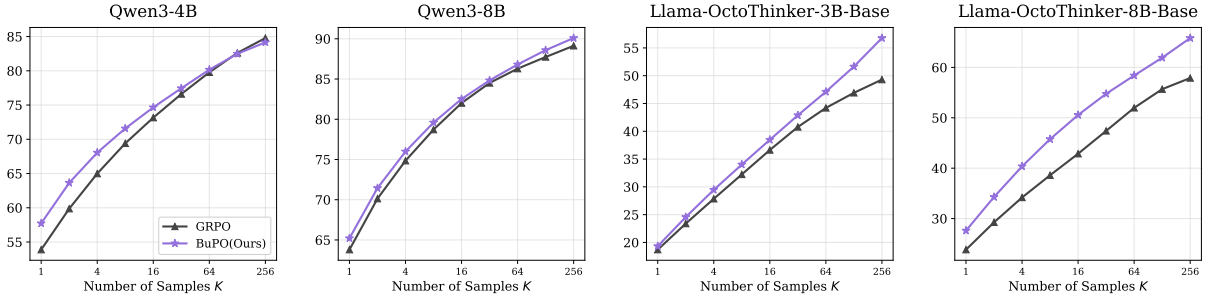


Figure 6: Average Pass@ K results on MATH500, AMC23, AIME24 and AIME25.

Qwen3-4B, the only exception occurs at $K = 256$. For all Llama models, BuPO achieves the best results across all K values, yielding gains of 7.48 points on Llama-OctoThinker-3B-Base and 7.93 points on Llama-OctoThinker-8B-Base in Pass@256. These results further indicate the IPA effectively enhances the reasoning capacity of π_θ .

5.2 Analysis

Training Dynamics of BuPO. We further visualize the training dynamics of BuPO. As shown in Figure 7, by aligning the internal layer policy at an early stage, all models exhibit enhanced entropy exploration initially. For Qwen models, optimizing layer 6 maintains stable exploration, consistent with the later layer-policy ablation. For Llama, we observe increased entropy during the bottom

alignment stage, indicating that feature refinement in lower layers effectively provides a larger exploration space for π_θ .

5.3 Ablation Study

Effect of Bottom Optimization Steps. We further study how the number of bottom optimization steps s_{inter} affects BuPO. As shown in Table 2, moderate internal policy alignment consistently improves over GRPO on Qwen3-4B. Specifically, setting $s_{\text{inter}} = 30$ improves the average score from 55.08 to 58.51, while increasing it to 50 or 70 steps causes a drop. These results are consistent with Figure 5: limited bottom optimization promotes internal policy alignment, whereas excessive alignment destabilizes the language model policy and harms final performance.

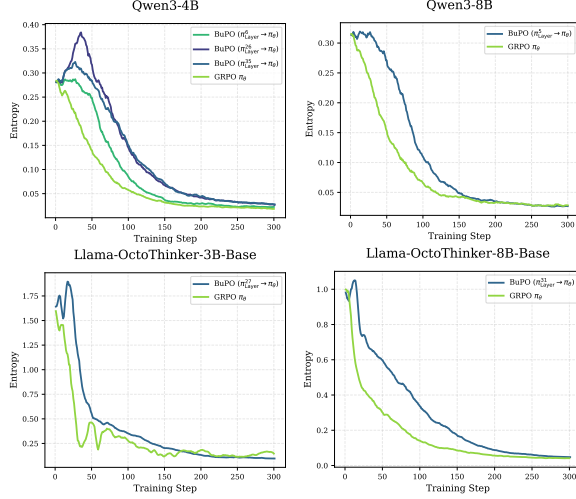


Figure 7: Entropy dynamics during training with GRPO and BuPO with different internal policy.

Effect of Optimized Internal Policy. We further study the effect of choosing different optimized internal policies π_{Layer}^l . As shown in Table 3, the best-performing layer varies across backbones: π_{Layer}^6 achieves the best performance on both Qwen3-4B and Qwen3-8B, while π_{Layer}^{26} and π_{Layer}^{30} perform best on OctoThinker-3B and OctoThinker-8B, respectively. Despite this variation, these best-performing layers share a common pattern in Figure 3: each corresponds to the last layer with a positive FFN exploration signal $\Delta H_{\text{FFN}}^l > 0$, i.e., the final layer before the FFN entropy dynamics shift from exploration to integration or convergence. For clarity, we refer to this entropy-indicated transition point as the **Boundary Layer**. Across all four backbones, optimizing the Boundary Layer achieves the best average performance, suggesting that the entropy-change pattern provides a useful practical signal for choosing the target internal policy. We analyze the computational cost of entropy-based inspection in Appendix A.6, showing that its overhead is small compared with downstream policy training.

6 Related Work

Reinforcement Learning with Verifiable Rewards. Recently, Reinforcement Learning with Verifiable Rewards (RLVR) has gained traction by leveraging rule-based rewards (Wang et al., 2025a; Liu et al., 2026b,a; Li et al., 2026; Yu et al., 2026; Zhou et al., 2026b,a; Liao et al., 2026; Liu et al., 2026c; Tan et al., 2026b; Zhu et al., 2026), a mechanism that has proven highly effective in fostering

Methods	AMC	MATH500	AIME24	AIME25	Average
Qwen3-4B					
GRPO	76.88	82.41	32.19	28.85	55.08
BuPO ($s_{\text{inter}} = 30$)	81.09	84.90	36.88	31.15	58.51
BuPO ($s_{\text{inter}} = 50$)	62.66	79.00	14.17	12.60	42.11
BuPO ($s_{\text{inter}} = 70$)	14.14	25.20	0.21	0.00	9.89

Table 2: Ablation of s_{inter} with fixed layer policy π_{Layer}^6 .

Layer Policy	AMC	MATH500	AIME24	AIME25	Average
Qwen3-4B					
Boundary Layer (π_{Layer}^6)	81.09	<u>84.90</u>	36.88	31.15	58.51
π_{Layer}^{26}	<u>81.56</u>	84.40	<u>35.42</u>	<u>30.52</u>	57.98
π_{Layer}^{35}	82.66	85.14	<u>35.42</u>	30.12	<u>58.34</u>
Qwen3-8B					
Boundary Layer (π_{Layer}^6)	89.22	87.76	54.06	34.38	66.36
π_{Layer}^{26}	75.55	84.40	32.60	21.67	53.56
π_{Layer}^{35}	<u>83.75</u>	<u>86.51</u>	<u>42.81</u>	<u>29.58</u>	<u>60.66</u>
OctoThinker-3B					
Boundary Layer (π_{Layer}^{26})	27.50	49.79	0.63	0.42	19.59
π_{Layer}^{16}	<u>16.17</u>	<u>11.40</u>	0.00	0.00	<u>6.89</u>
π_{Layer}^6	7.80	9.60	<u>0.21</u>	<u>0.10</u>	4.43
OctoThinker-8B					
Boundary Layer (π_{Layer}^{30})	37.66	62.05	4.69	6.77	27.79
π_{Layer}^{16}	3.91	<u>12.20</u>	0.00	0.00	4.03
π_{Layer}^9	<u>7.58</u>	9.40	0.00	0.10	<u>4.27</u>

Table 3: Ablation of the optimized internal policy π_{Layer}^l . The **Boundary Layer** denotes the target layer selected for BuPO, corresponding to the last layer with positive FFN entropy change.

the reasoning capabilities of LLMs (Guo et al., 2025; Yang et al., 2025; Team et al., 2025; Rank et al., 2026; Tan et al., 2026a). In this work, we shift the focus to internal policies and propose bottom-up policy optimization, which directly optimizes internal layer policies in early training stages. This targeted optimization refines internal reasoning representations and ultimately leads to improved performance.

Interpretability of LLMs. Interpretability studies have opened LLM black boxes by analyzing how models reason, store knowledge, and route information internally (Tan et al., 2025; Gupta et al., 2025; Hu et al., 2025c; Zhao et al., 2025, 2026), especially through self-attention (Zhou et al., 2025; Jin et al., 2025) and FFN (Dai et al., 2022; Meng et al., 2022). In this work, we conduct a systematic analysis of hidden states from a policy-centric perspective, revealing structured internal reasoning patterns that motivate internal policy alignment which refines intermediate features before full policy optimization.

7 Conclusion

In this paper, we decompose a language model policy into internal layer and modular policies, revealing systematic reasoning patterns through en-

tropy analysis. We observe a broad transition from high-entropy exploration in early layers to deterministic convergence in higher layers, with Qwen models showing a progressive reasoning structure contrasting with the abrupt convergence observed in Llama. Motivated by these findings, we propose Bottom-up Policy Optimization (BuPO), an RL paradigm that aligns internal layer policies during early training. Extensive experiments on complex reasoning benchmarks demonstrate its effectiveness. Further analysis shows that internal policy alignment refines foundational reasoning features, thereby improving the reasoning capacity of the overall policy.

Limitations

While BuPO demonstrates consistent improvements by leveraging internal policy optimization, several practical limitations offer avenues for future research. (1) Similar to existing RLVR methods, BuPO requires substantial computational resources for rollout generation, reward evaluation, and policy optimization. Our experiments were conducted on a single node with 8 NVIDIA A100 GPUs. Such training requirements may limit accessibility for researchers with constrained computational budgets, and improving the efficiency of RL training remains an important direction for the community. (2) BuPO is studied in reasoning-oriented settings, where models often generate long reasoning trajectories before producing final answers. This increases both training and evaluation costs, especially when evaluating reasoning performance with large-sample metrics such as Pass@ K with 300 sampled responses. Overall, we hope BuPO inspires the design of more effective RL algorithms that make use of internal model structures, and encourages closer connections between reinforcement learning and interpretability for reasoning language models.

Ethical considerations

Our approach does not introduce ethical concerns. The datasets we used are public, and there are no privacy issues.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62376270) and the Independent Research Project of the Key Laboratory of Cognition and Decision Intelligence for

Complex Systems.

References

- Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. 2025. [The unreasonable effectiveness of entropy minimization in LLM reasoning](#). In *Proceedings of NeurIPS*.
- Arash Ahmadian, Chris Cremer, Matthias Gall , Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet  st n, and Sara Hooker. 2024. [Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs](#). In *Proceedings of ACL*, pages 12248–12267.
- Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. [Eliciting latent predictions from transformers with the tuned lens](#). *arXiv preprint arXiv:2303.08112*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Proceedings of NeurIPS*, volume 33, pages 1877–1901.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. [Evaluating large language models trained on code](#). *arXiv preprint arXiv:2107.03374*.
- Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. 2025. [Reasoning with exploration: An entropy perspective](#). *arXiv preprint arXiv:2506.14758*.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have solved question answering? try arc, the ai2 reasoning challenge](#). *arXiv preprint arXiv:1803.05457*.
- Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, and 1 others. 2025. [The entropy mechanism of reinforcement learning for reasoning language models](#). *arXiv preprint arXiv:2505.22617*.
- Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. [Knowledge neurons in pretrained transformers](#). In *Proceedings of ACL*, pages 8493–8502.
- Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. [Transformer feed-forward layers are key-value memories](#). In *Proceedings of EMNLP*, pages 5484–5495.

- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. [Deepseek-r1 incentivizes reasoning in llms through reinforcement learning](#). *Nature*, 645(8081):633–638.
- Akshat Gupta, Jay Yeung, Gopala Anumanchipalli, and Anna Ivanova. 2025. [How do llms use their depth?](#) *arXiv preprint arXiv:2510.18871*.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. [OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems](#). In *Proceedings of ACL*, pages 3828–3850.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. [Deep residual learning for image recognition](#). In *Proceedings of CVPR*, pages 770–778.
- Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, and 1 others. 2025. [Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning](#). *arXiv preprint arXiv:2504.11456*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. [Measuring mathematical problem solving with the MATH dataset](#). In *Proceedings of NeurIPS Datasets and Benchmarks Track*.
- Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. 2025a. [Reinforce++: Stabilizing critic-free policy optimization with global advantage normalization](#). *arXiv preprint arXiv:2501.03262*.
- Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. 2025b. [Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model](#). *arXiv preprint arXiv:2503.24290*.
- Yi Hu, Cai Zhou, and Muhan Zhang. 2025c. [What affects the effective depth of large language models?](#) In *Mechanistic Interpretability Workshop at NeurIPS 2025*.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Barney, and 1 others. 2024. [Openai o1 system card](#). *arXiv preprint arXiv:2412.16720*.
- Mingyu Jin, Kai Mei, Wujiang Xu, Mingjie Sun, Ruixiang Tang, Mengnan Du, Zirui Liu, and Yongfeng Zhang. 2025. [Massive values in self-attention modules are the key to contextual knowledge understanding](#). In *Proceedings of ICML*.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with pagedattention](#). In *Proceedings of SOSP*, pages 611–626.
- Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and 1 others. 2022. Solving quantitative reasoning problems with language models. In *Proceedings of NeurIPS*, volume 35, pages 3843–3857.
- Yang Li, Zhichen Dong, Yuhan Sun, Weixun Wang, Shaopan Xiong, Yijia Luo, Jiashun Liu, Han Lu, Jiamang Wang, Wenbo Su, and 1 others. 2025. [Attention illuminates llm reasoning: The preplan-and-anchor rhythm enables fine-grained policy optimization](#). *arXiv preprint arXiv:2510.13554*.
- Zhaoyi Li, Deyang Kong, Yuan Wei, Evan Yang, Rannan Shen, Mahardika Krisna Ihsani, Ming Yang, Wei Zhang, Chuan Hao, Jian Yang, and 1 others. 2026. [Every coin has two sides: On the dual nature of generalization in on-policy distillation of large language models](#). *arXiv preprint arXiv:2608.16647*.
- Huanxuan Liao, Zhongtao Jiang, Yupu Hao, Yuqiao Tan, Shizhu He, Ben Wang, Jun Zhao, Kun Xu, and Kang Liu. 2026. [Resadapt: Adaptive resolution for efficient multimodal reasoning](#). *arXiv preprint arXiv:2603.28610*.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. [Let’s verify step by step](#). In *International Conference on Learning Representations*, volume 2024, pages 39578–39601.
- Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermy, Andy Jones, and 8 others. 2025. [On the biology of a large language model](#). *Transformer Circuits Thread*.
- Runze Liu, Jiakang Wang, Yuling Shi, Zhihui Xie, Chenxin An, Kaiyan Zhang, Jian Zhao, Xiaodong Gu, Lei Lin, Wenping Hu, and 1 others. 2025. [Attention as a compass: Efficient exploration for process-supervised rl in reasoning models](#). *arXiv preprint arXiv:2509.26628*.
- Weiting Liu, Jieyi Bi, Wanqi Zhou, Jianfeng Feng, Yining Ma, Ai Han, and Wenlian Lu. 2026a. [Toolanchor: Anchoring counterfactual context to boost agentic tool-use capability](#). *arXiv preprint arXiv:2607.14145*.

- Weiting Liu, Han Wu, Yufei Kuang, Xiongwei Han, Tao Zhong, Jianfeng Feng, and Wenlian Lu. 2026b. [Automated optimization modeling via a localizable error-driven perspective](#). *arXiv preprint arXiv:2602.11164*.
- Zirui Liu, Jie Ouyang, Qi Liu, Xianquan Wang, Jiayu Liu, Tingyue Pan, Qingchuan Li, Jing Sha, Zhenya Huang, Shijin Wang, and 1 others. 2026c. [Socraticpo: Policy optimization via interactive guidance](#). *arXiv preprint arXiv:2606.09887*.
- MAA. 2023. [American mathematics contest 12 \(amc 12\)](#).
- MAA. 2024. [American invitational mathematics examination \(aime\)](#).
- MAA. 2025. [American invitational mathematics examination \(aime\)](#).
- Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. [Locating and editing factual associations in GPT](#). In *Proceedings of NeurIPS*.
- Meta AI. 2024. [Llama 3.2: Revolutionizing edge ai and vision with open, customizable models](#). Accessed: 2025-11-03.
- nostalgebraist. 2020. [interpreting gpt: the logit lens](#). LessWrong.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Proceedings of NeurIPS*, volume 35, pages 27730–27744.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Proceedings of NeurIPS*, volume 36, pages 53728–53741.
- Ben Rank, Hardik Bhatnagar, Ameya Prabhu, Shira Eisenberg, Karina Nguyen, Matthias Bethge, and Maksym Andriushchenko. 2026. [Posttrainbench: Can llm agents automate llm post-training?](#) *arXiv preprint arXiv:2603.08640*.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. [Gpqa: A graduate-level google-proof q&a benchmark](#). In *First Conference on Language Modeling*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *arXiv preprint arXiv:2402.03300*.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. [Hybridflow: A flexible and efficient rlhf framework](#). In *Proceedings of EuroSys*, pages 1279–1297.
- Richard S Sutton, Andrew G Barto, and 1 others. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and 1 others. 2023. [Challenging big-bench tasks and whether chain-of-thought can solve them](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051.
- Yuqiao Tan, Shizhu He, Kang Liu, and Jun Zhao. 2025. [Neural incompatibility: The unbridgeable gap of cross-scale parametric knowledge transfer in large language models](#). In *Proceedings of ACL*, pages 21586–21601.
- Yuqiao Tan, Minzheng Wang, Shizhu He, Jun Zhao, and Kang Liu. 2026a. [PostTrainBench⁰: Can llm agents automate llm post-training without gradients?](#) Research blog.
- Yuqiao Tan, Minzheng Wang, Bo Liu, Zichen Liu, Tian Liang, Shizhu He, Jun Zhao, and Kang Liu. 2026b. [From \$p\(y|x\)\$ to \$p\(y\)\$: Investigating reinforcement learning in pre-train space](#). *arXiv preprint arXiv:2604.14142*.
- Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. [Kimi k1. 5: Scaling reinforcement learning with llms](#). *arXiv preprint arXiv:2501.12599*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Proceedings of NeurIPS*, page 6000–6010.
- Minzheng Wang, Yongbin Li, Haobo Wang, Xinghua Zhang, Nan Xu, Bingli Wu, Fei Huang, Haiyang Yu, and Wenji Mao. 2025a. [Adaptive thinking via mode policy optimization for social language agents](#). *arXiv preprint arXiv:2505.02156*.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, and 1 others. 2024. [Mmlu-pro: A more robust and challenging multi-task language understanding benchmark](#). *Advances in Neural Information Processing Systems*, 37:95266–95290.
- Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. 2025b. [Octothinker: Mid-training incentivizes reinforcement learning scaling](#). *arXiv preprint arXiv:2506.20512*.

- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, and 1 others. 2024. [Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement](#). *arXiv preprint arXiv:2409.12122*.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gao-hong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, and 17 others. 2025. [DAPO: An open-source LLM reinforcement learning system at scale](#). In *Proceedings of NeurIPS*.
- Yi Yu, Junzhuo Ma, Chenghuang Shen, Xingyan Liu, Jing Gu, Hangyi Sun, Guangquan Hu, Jianfeng Liu, Weiting Liu, Mingyue Pu, and 1 others. 2026. [Lightweight adaptation for llm-based technical service agent: Latent logic augmentation and robust noise reduction](#). *arXiv preprint arXiv:2603.18074*.
- Zeping Yu and Sophia Ananiadou. 2024. [Neuron-level knowledge attribution in large language models](#). In *Proceedings of EMNLP*, pages 3267–3280.
- Biao Zhang and Rico Sennrich. 2019. [Root mean square layer normalization](#). In *Proceedings of NeurIPS*, volume 32.
- Honglin Zhang, Qianyu Hao, Fengli Xu, and Yong Li. 2025. [Reinforcement learning fine-tuning enhances activation intensity and diversity in the internal circuitry of llms](#). *arXiv preprint arXiv:2509.21044*.
- Chengfeng Zhao, Shizhu He, Shanshan Jiang, Bin Dong, Jun Zhao, and Kang Liu. 2025. [Why and how llms benefit from knowledge introspection in common-sense reasoning](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6905–6920.
- Chengfeng Zhao, Yuqiao Tan, Shizhu He, Yequan Wang, Jun Zhao, and Kang Liu. 2026. [Neural procedural memory: Empowering llm agents with implicit activation steering](#). *arXiv preprint arXiv:2606.29824*.
- Yixiao Zhou, Dongzhou Cheng, Zhiliang Wu, Yi Yang, Yu Cheng, and Hehe Fan. 2026a. One refiner to unlock them all: Inference-time reasoning elicitation via reinforcement query refinement. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 38957–38978.
- Yixiao Zhou, Yang Li, Dongzhou Cheng, Hehe Fan, and Yu Cheng. 2026b. Look inward to explore outward: Learning temperature policy from llm internal states via hierarchical rl. *arXiv preprint arXiv:2602.13035*.
- Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. 2025. [On the role of attention heads in large language model safety](#). In *Proceedings of ICLR*.
- Chunzheng Zhu, Yangfang Lin, Shen Chen, Yijun Wang, and Jianxin Lin. 2026. [Medeyes: Learning dynamic visual focus for medical progressive diagnosis](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 13916–13924.

A Detailed Experiment Settings

A.1 The Models for Experiments

We summarize all models used in our analysis and experiments in Table 4. These models are categorized into three types: Mix, Base, and Instruct.

A.2 The Template for Experiments

We adopt the following template for all experiments involving Qwen models, building upon the Qwen-Math template used for Qwen2.5 (Yang et al., 2024) and the Qwen-NoThinking template for Qwen3.

Qwen-Math Template

```
<|im_start|>system
Please reason step by step, and put your
final answer within \boxed{ }. <|im_end|>
<|im_start|>user
{problem}
<|im_end|>
<|im_start|>assistant
```

Figure 8: Prompt template for Qwen-math.

Qwen3-NoThinking Template

```
<|im_start|>system
Please reason step by step, and put your fi-
nal answer within \boxed{ }. <|im_end|>
<|im_start|>user
{problem}
<|im_end|>
<|im_start|>assistant
<think>

</think>
```

Figure 9: Prompt template for Qwen3 NoThinking mode.

For training the Llama-OctoThinker models, we adopt the original prompt in Wang et al. (2025b) to ensure performance.

A.3 Implementation of Internal Policy Entropy Analysis

In this section, we detail the implementation of internal policy entropy analysis. Our primary objective is to extract internal hidden states during the

OctoThinker Template

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. User: You must put your answer inside \boxed{ } and Your final answer will be extracted automatically by the \boxed{ } tag. {problem} Assistant:

Figure 10: Prompt template for OctoThinker.

forward pass. In the main experiments, we evaluate model-generated responses on the MATH test set (Hendrycks et al., 2021). Entropy is computed at the token level for each layer and module, and then averaged over all generated tokens. We find that the intrinsic reasoning patterns remain stable across different tasks, e.g., commonsense question answering (Rein et al., 2024).

The computation of internal policy entropy is illustrated in the pseudo-code below. We abstract the entropy computation as a function $\mathcal{H}(\cdot)$. Accordingly, the entropy change of the internal layer policy is defined as:

$$\Delta H_{\text{Layer}}^l = \mathcal{H}(\mathbf{H}^l) - \mathcal{H}(\mathbf{H}^{l-1}). \quad (11)$$

For the two core Transformer submodules, the entropy changes are computed separately. Specifically, for the self-attention module, we define:

$$\Delta H_{\text{ATTN}}^l = \mathcal{H}(\mathbf{A}^l) - \mathcal{H}(\text{LN}(\mathbf{H}^{(2l-2)})), \quad (12)$$

and for the feed-forward network (FFN), we define:

$$\Delta H_{\text{FFN}}^l = \mathcal{H}(\mathbf{F}^l) - \mathcal{H}(\text{LN}(\mathbf{H}^{(2l-1)})). \quad (13)$$

These definitions allow us to quantify how each layer and submodule contributes to the evolution of the internal policy entropy.

Discussion: Comparison to the Logit Lens. Notably, our definition of internal policy differs from the logit-lens (nostalgebraist, 2020), particularly in how layer normalization (LN) is handled. To clarify this distinction, we provide a systematic comparison between the two formulations in Table 6.

Model	Huggingface	Type	Layers
Qwen3-4B	https://huggingface.co/Qwen/Qwen3-4B	Mix	36
Qwen3-8B	https://huggingface.co/Qwen/Qwen3-8B	Mix	36
Qwen3-14B	https://huggingface.co/Qwen/Qwen3-14B	Mix	40
Qwen3-4B-Base	https://huggingface.co/Qwen/Qwen3-4B-Base	Base	36
Qwen2.5-Math-7B	https://huggingface.co/Qwen/Qwen2.5-Math-7B	Base	28
Qwen3-4B-Instruct-2507	https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507	Instruct	36
Llama-3.2-3B-Instruct	https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct	Instruct	28
Llama-3.1-8B-Instruct	https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct	Instruct	32
Llama-OctoThinker-3B-Base	https://huggingface.co/OctoThinker/OctoThinker-3B-Long-Base	Base	28
Llama-OctoThinker-8B-Base	https://huggingface.co/OctoThinker/OctoThinker-8B-Long-Base	Base	32
DeepSeek-Math-7B-Base	https://huggingface.co/deepseek-ai/deepseek-math-7b-base	Base	30
DeepSeek-R1-Distill-Qwen-7B	https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B	Instruct	28

Table 4: Detailed information about the selected models is provided. "Mix" refers to models that support both thinking and non-thinking modes. "Base" denotes the pre-trained model only. "Instruct" indicates models that undergo further fine-tuning based on the Base model to enhance instruction-following capabilities.

Calculation of Internal Policy Entropy (PyTorch Implementation)

```
# Get layer hidden states by register hook
hidden_state = get_from_hook()

# Compute logits in the same way as in the original forward pass
logits = self.model.lm_head(hidden_state)

# Apply softmax for normalization
probs = torch.softmax(logits, dim=-1)

# Apply log_softmax for speedy computation
log_probs = torch.log_softmax(logits, dim=-1)

# Calculate internal layer policy
entropies = -(probs * log_probs).sum(dim=-1)
```

Figure 11: The pseudocode of internal policy entropy calculation.

Our definition adopts a policy-centric perspective, treating the internal hidden states as explicitly policy distribution. In contrast, the logit-lens is primarily designed to project hidden states into the discrete vocabulary space in order to inspect the most likely output tokens at intermediate layers. We intentionally omit LN based on empirical considerations: in our analysis experiments, incorporating LN leads to less stable entropy dynamics and weaker interpretability than those shown in Figures 2 and 3.

Extensive analyses and experiments with BuPO further demonstrate that our formulation of internal policy serves as a robust interpretability tool for uncovering internal reasoning mechanisms in LLMs. As shown in Table 5, we also provide an ablation on layer normalization, finding that removing LN yields consistently better performance, which validates our design choice of omitting it in BuPO.

Methods	AMC	MATH	AIME24	AIME25
Qwen3-4B				
BuPO-wo-LN (Ours)	81.09	84.90	36.88	31.15
BuPO-w-LN	80.08	84.80	32.29	30.00
Qwen3-8B				
BuPO-wo-LN (Ours)	89.22	87.76	54.06	34.38
BuPO-w-LN	86.41	87.40	47.81	34.21

Table 5: Ablation on layer normalization in BuPO.

A.4 Implementation of Internal Policy Alignment

In internal policy alignment, namely IPA, at each optimization step, we select a specific internal layer l and optimize its internal policy π_{Layer}^l , defined as $\pi_{\text{Layer}}^l = \text{softmax}(\mathbf{H}^l \mathbf{E}_u^T)$, where $\mathbf{H}^l$ is the hidden states at layer l . The gradient flow for internal policy optimization is determined by the residual structure of the Transformer, where the hidden state $\mathbf{H}^l$ is a function of all parameters from layer 1 to l ,

	Logit Lens	Ours
Perspective	Discrete Token	Policy Distribution
Definition	$\text{LN}(\mathbf{H}^l)\mathbf{E}_u^T$	$\text{softmax}(\mathbf{H}^l\mathbf{E}_u^T)$
Trainable	✗	✓

Table 6: Comparison between the logit lens and our definition

but is independent of parameters in higher layers.

Formally, for any parameter θ_k in layer k , the gradient of the IPA loss with respect to θ_k can be expressed using the chain rule:

$$\frac{\partial J_{\text{IPA}}(\pi_{\text{Layer}}^l)}{\partial \theta_k} = \frac{\partial J_{\text{IPA}}}{\partial \pi_{\text{Layer}}^l} \cdot \frac{\partial \pi_{\text{Layer}}^l}{\partial \mathbf{H}^l} \cdot \frac{\partial \mathbf{H}^l}{\partial \theta_k} \quad (14)$$

Due to the residual connections, $\frac{\partial \mathbf{H}^l}{\partial \theta_k} \neq 0$ only when $k \leq l$, and is zero otherwise. Thus, the gradients for different layers can be summarized as:

$$\frac{\partial J_{\text{IPA}}(\pi_{\text{Layer}}^l)}{\partial \theta_k} = \begin{cases} \frac{\partial J_{\text{IPA}}}{\partial \pi_{\text{Layer}}^l} \cdot \frac{\partial \pi_{\text{Layer}}^l}{\partial \mathbf{H}^l} \cdot \frac{\partial \mathbf{H}^l}{\partial \theta_k}, & \text{if } k \leq l \\ 0, & \text{if } k > l \end{cases} \quad (15)$$

This means that, during IPA optimization for layer l , only the parameters of layers 0 through l and unembedding matrix $\mathbf{E}_u$ are updated, while all higher layers ($k > l$) remain unaffected. This targeted gradient flow ensures that internal policy optimization provides direct supervision to the selected layer and all lower layers, strengthening foundational reasoning capabilities without interfering with higher-level representations.

Training Setup. Based on the findings in Section 3, we select Qwen3-4B (Yang et al., 2025) in non-thinking mode for investigation. For the training set, we randomly sample 5k entries from *DeepMath-103k* (He et al., 2025). We train the models using the verl framework (Sheng et al., 2025). The prompt batch size is 128, with 8 roll-outs generated per prompt. The sampling temperature during training is set to 1.0, and the maximum context length is set to 9,216 tokens. We update the model with a mini-batch size of 32 for 300 steps and a learning rate of 1e-6. These stage-wise FFN entropy patterns suggest that transition points between exploration and later integration/convergence may provide informative candidates for internal policy optimization. For Qwen3-4B, the region

boundaries identified in Section 3.2 lie at layers 6 and 26. Additionally, we focus on aligning the internal policy of the penultimate layer, which is critical for the final prediction. Accordingly, we compare the internal policies π_{Layer}^6 , π_{Layer}^{26} , and π_{Layer}^{35} optimized via IPA, against the overall policy π_θ optimized with GRPO (He et al., 2025).

A.5 Implementation of Main Experiments

Detailed Training Setup. Specifically, we focus on the Qwen3 series, using Qwen3-4B and Qwen3-8B (Yang et al., 2025). Meanwhile, we select Llama-OctoThinker-3B-Base and Llama-OctoThinker-8B-Base from the Llama series, as these models demonstrate improved training behavior after mid-training based on Llama-3.2-Base (Wang et al., 2025b). We set $s_{\text{inter}} = 30$ for Qwen3-4B and $s_{\text{inter}} = 20$ for the other LLMs. We implement GRPO and other baseline algorithms using the verl framework (Sheng et al., 2025). Across all algorithms and model variants, we adopt a unified set of hyperparameters, as reported in Table 8, and do not employ entropy regularization or KL-based losses. For PPO, the critic network is trained separately with a learning rate of 1×10^{-5} . All experimental results are averaged over three random seeds.

Evaluation Setup. We use vLLM (Kwon et al., 2023) with temperature 1.0 and top_p 1.0. This metric is defined as $\text{Pass}@K := \mathbb{E}_{x \sim \mathcal{D}} [1 - \binom{n-c}{K} / \binom{n}{K}]$, where c denotes the number of correct completions out of n generated responses. To reduce evaluation variance on those datasets, we set $n = 300$.

A.6 The Cost of Internal Policy Analysis

The internal policy analysis is a lightweight one-time diagnostic used to identify the layer for bottom-up policy optimization. All experiments are conducted on a single node with 8 NVIDIA A100 GPUs. As shown in Figure 11, for each sample, the analysis only requires a standard forward pass: we register hooks to cache intermediate hidden states, project them with the shared unembedding matrix, and compute token-level entropy. Therefore, it introduces no backward pass, policy update or reference-model pass.

Time Cost. Taking Qwen3-4B as an example, the full RL training takes 13.22 hours, i.e., about 47,592 seconds. In contrast, the internal policy entropy analysis takes only 209.25 seconds. This cor-

Metric	RL Training	Analysis	Ratio
Total Time	13.22h ($\sim 47,592$ s)	209.25s	$\sim 1:227.4$
Total FLOPs	12288 S TFLOPs	17.6 S TFLOPs	$\sim 1:698.2$

Table 7: Cost comparison between RL training and internal policy analysis on Qwen3-4B.

Hyperparameter	Value
Optimizer	AdamW
Policy learning rate	$1e^{-6}$
Critic learning rate	$1e^{-5}$ (for PPO)
Training batch size	128 prompts
Samples per prompt	8
Mini-batch size	32 prompts
Policy updates per rollout	16
Max prompt length	1024 tokens
Max response length	7168 tokens (Qwen) / 3072 (Llama)
Rollout temperature	1.0
Clip range ϵ	0.2

Table 8: RL Hyperparameters

responds to about 0.44% of the RL training time, or an analysis-to-training ratio of approximately 1 : 227.4. Such a small one-time cost is sufficient to recover stable internal reasoning patterns for selecting the optimized internal layer.

FLOPs Cost. Let P denote the number of model parameters, B the number of analyzed samples, and S the average token count per sample. A typical RL step contains generation, policy forward, backward/update, and reference passes, costing roughly $10PS$ FLOPs per sample. For Qwen3-4B, with 300 steps and 1,024 samples per step, the total training cost is approximately $300 \times 1024 \times 10 \times 4B \times S \approx 12288 S$ TFLOPs. By comparison, the analysis requires only a single forward pass, about $2PBS$, plus entropy computation dominated by LM-head projections across internal layers. The total cost is approximately 17.6 S TFLOPs, yielding an analysis-to-training FLOPs ratio of about 1 : 698.2.

Moreover, the analysis can be directly performed on the baseline model rollouts at step 0, requiring no additional rollout collection. These results show that internal policy analysis is substantially cheaper than downstream RL training while providing the necessary signal for practical layer selection.

B Extended Experiment Results

B.1 Out-of-Domain Generalization

To evaluate the generalization ability of BuPO, we extend our evaluation to broader benchmarks spanning advanced math (OlympiadBench (He et al., 2024), Minerva (Lewkowycz et al., 2022)), general reasoning (MMLU-Pro (Wang et al., 2024), GPQA-D (Rein et al., 2024), ARC-C (Clark et al., 2018), BBH (Suzgun et al., 2023)), and coding (HumanEval (Chen et al., 2021)). As shown in Table 9, BuPO consistently outperforms both the vanilla baseline and GRPO across all benchmarks on both model scales. BuPO achieves average gains of +1.83 and +0.62 points over GRPO on Qwen3-4B and Qwen3-8B, respectively, with particularly notable improvements on GPQA-D (+4.10 for 4B, +1.50 for 8B) and HumanEval (+3.60 for 4B, +1.83 for 8B). These consistent gains across diverse tasks suggest that optimizing internal layer policies encourages the model to develop more transferable reasoning capabilities rather than overfitting to training task patterns.

B.2 Comparison with Conventional Layer-aware Optimization

To examine whether BuPO specifically benefits from internal policy optimization, rather than merely from emphasizing lower layers, we compare it with two conventional layer-aware GRPO variants. The first baseline freezes layers up to the selected layer l , while the second applies a smaller learning rate ($0.1 \times$) to layers above l . As shown in Table 11, BuPO consistently outperforms both alternatives on Qwen3-4B and OctoThinker-3B. These results show that simply freezing layers or reweighting learning rates is insufficient to match BuPO. In contrast, optimizing the internal policy through IPA provides a more effective training signal, confirming the effectiveness of BuPO.

B.3 Multi-seed Robustness

To address whether the reported improvements depend on a single random seed, we rerun both GRPO and BuPO with three random seeds under the same training setup. Table 10 reports the mean and standard deviation across seeds. BuPO outperforms GRPO on most benchmarks. Averaged over the four benchmarks, BuPO improves over GRPO by +3.77, +2.53, +1.95, and +3.62 points on Qwen3-4B, Qwen3-8B, OctoThinker-3B, and OctoThinker-8B, respectively.

Methods	Olympiad	Minerva	MMLU-Pro	GPQA-D	ARC-C	BBH	HumanEval	Avg
Qwen3-4B								
Vanilla	32.27	23.99	60.50	45.50	87.40	75.50	78.70	57.69
GRPO	36.15	28.77	61.30	43.90	87.10	75.00	79.30	58.79
BuPO (Ours)	37.52	29.89	61.70	48.00	87.50	76.80	82.90	60.62
Δ vs. GRPO	+1.37	+1.12	+0.40	+4.10	+0.40	+1.80	+3.60	+1.83
Qwen3-8B								
Vanilla	32.51	22.45	63.40	51.00	90.80	79.60	83.54	60.47
GRPO	40.31	31.55	66.60	52.00	90.87	81.55	84.15	63.86
BuPO (Ours)	40.63	31.92	66.80	53.50	91.00	81.50	85.98	64.48
Δ vs. GRPO	+0.32	+0.37	+0.20	+1.50	+0.13	-0.05	+1.83	+0.62

Table 9: Out-of-domain evaluation results across general reasoning and coding benchmarks. **Bold** denotes the best.

Methods	AMC	MATH	AIME24	AIME25
Qwen3-4B				
BuPO	80.78$\pm$0.41	85.19$\pm$0.26	36.77$\pm$0.11	30.99$\pm$0.63
GRPO	76.67 $\pm$ 0.36	83.12 $\pm$ 0.62	32.15 $\pm$ 0.47	26.70 $\pm$ 1.93
Qwen3-8B				
BuPO	89.06$\pm$0.72	88.00 $\pm$ 0.22	53.85$\pm$0.18	34.90$\pm$0.48
GRPO	86.31 $\pm$ 0.39	88.02$\pm$0.06	48.19 $\pm$ 1.36	33.19 $\pm$ 0.30
OctoThinker-3B				
BuPO	27.71$\pm$0.24	49.88$\pm$0.17	0.66$\pm$0.06	3.86$\pm$0.08
GRPO	27.29 $\pm$ 0.18	46.42 $\pm$ 0.31	0.45 $\pm$ 0.16	0.17 $\pm$ 0.06
OctoThinker-8B				
BuPO	38.08$\pm$0.59	62.02$\pm$0.05	4.76$\pm$0.12	6.67$\pm$0.11
GRPO	35.16 $\pm$ 0.41	57.12 $\pm$ 0.21	2.64 $\pm$ 0.24	2.15 $\pm$ 0.06

Table 10: Multi-seed results of BuPO and GRPO. We report mean $\pm$ standard deviation over three random seeds.

Methods	AMC	MATH	AIME24	AIME25	Avg.
Qwen3-4B					
BuPO	81.09	84.90	36.88	31.15	58.51
Frozen layers	78.91	83.60	31.04	29.06	55.65
Layerwise LR	80.86	84.00	36.88	30.73	58.12
OctoThinker-3B					
BuPO	27.50	49.79	0.63	0.42	19.59
Frozen layers	23.67	46.40	0.31	0.42	17.70
Layerwise LR	26.17	44.80	0.63	0.21	17.95

Table 11: Comparison with conventional layer-aware GRPO alternatives.

The standard deviations are generally small relative to the performance gaps, indicating that the improvements are stable across random seeds. These results show that the gains of BuPO are not an artifact of a single run and further support the reliability of the main results.

B.4 Internal Policy Entropy Dynamics for More Models

In this section, we present additional preliminary analyses of internal policy entropy dynamics across a broader set of models. Specifically, we examine different variants of the same backbone, including

Base, Instruct, and Mix versions, as well as models trained with supervised fine-tuning (SFT) and reinforcement learning (RL). In addition, we include the DeepSeek-Math model (He et al., 2025) to further enrich the comparative analysis.

Further Training Has Limited Impact on Internal Reasoning Patterns. We further analyze models that undergo additional training beyond standard pre-training. Specifically, we include DeepSeek-R1-Distill-Qwen-7B, which is further trained from Qwen2.5-Math-7B using distilled responses from Guo et al. (2025) with SFT, as well as Llama-OctoThinker-3B-Base and Llama-OctoThinker-8B-Base, which are obtained via continued pre-training (i.e., mid-training) based on Llama-3.2-3B-Base and Llama-3.2-8B-Base, respectively (Wang et al., 2025b). Moreover, Qwen3-4B and Qwen3-4B-Base also show consistent pattern after post-training with RL. Notably, these additional training procedures exhibit only marginal influence on the internal reasoning patterns.

In summary, reinforcement learning, supervised fine-tuning, and mid-training do not substantially alter the model’s internal reasoning mechanisms, suggesting that these intrinsic patterns are primarily determined by the model architecture and initial pre-training. Understanding how such patterns emerge remains an important direction for future work.

Same Series of Models Exhibit Consistent Structures. After analyzing in all analysis plots across models, we find that models from the same series show consistent structures. For instance, all Qwen3 series models show progressive internal reasoning patterns including Qwen3-4B, Qwen3-8B and Qwen3-14B, also with other base or instruct version. Also, Llama-3.1 and Llama-3.2 show intra-

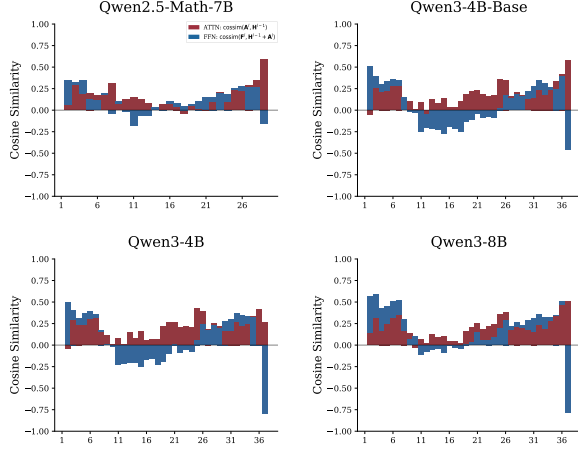


Figure 12: Residual cosine similarity across different Qwen models.

difference and inter-consistency.

Entropy Dynamics of DeepSeek-Math. We further analyze the internal reasoning mechanisms of DeepSeek-Math-7B-Base to provide a more comprehensive comparison. As illustrated in Figure 13, DeepSeek-Math-7B-Base exhibits a markedly different entropy dynamics: the overall internal policy entropy decreases substantially and converges primarily in the middle layers. This phenomenon is primarily driven by the consistently negative entropy change in the FFN module, i.e., $\Delta H_{\text{FFN}}^l < 0$, as illustrated in Figure 14, particularly in the middle layers.

Based on this observation, we infer that both Qwen and DeepSeek-Math demonstrate strong capability in knowledge absorption during post-training, indicating that convergence behavior in internal reasoning plays a critical role in effective learning, in contrast to Llama. Moreover, the generation search space of DeepSeek-Math appears more constrained than that of Qwen, particularly the Qwen3 series, suggesting reduced exploration capacity. We hypothesize that such internal reasoning patterns significantly influence the effectiveness of further training, pointing to promising directions for architectural design and optimization of foundation models.

B.5 How do Internal modules influence the residual stream?

To further understand how internal modules shape the residual stream in Qwen models with a progressive reasoning pattern, we analyze residual cosine similarity, which quantifies how each module writes to the residual pathway (Hu et al., 2025c).

For a given layer l , we compute $\text{cossim}(\mathbf{A}^l, \mathbf{H}^{l-1})$ for self-attention and $\text{cossim}(\mathbf{F}^l, \mathbf{H}^{l-1} + \mathbf{A}^l)$ for the FFN. A cosine similarity near zero indicates writing new, orthogonal features; negative values indicate feature suppression; and positive values indicate amplification of existing features.

As shown in Figure 12, the Qwen models largely follow the entropy dynamics discussed earlier, while exhibiting clear inter-generation differences. For Qwen3, self-attention consistently amplifies the residual stream, in line with its positive entropy change and expanded exploration behavior. In contrast, Qwen2.5 shows noticeably weaker attention write-in strength, with reduced cosine similarity magnitudes, consistent with its negative entropy change in self-attention.

The FFN modulates the residual stream in a stage-dependent manner across all models: in lower layers, it injects largely orthogonal features to support exploration; in middle layers, it suppresses vague signals while integrating parametric knowledge in FFN, corresponding to the Integration stage; and in upper layers, it amplifies and integrates features to drive convergence. Across all models, the final layer exhibits a sharp directional shift, underscoring its critical role in final prediction (Gupta et al., 2025; Agarwal et al., 2025).

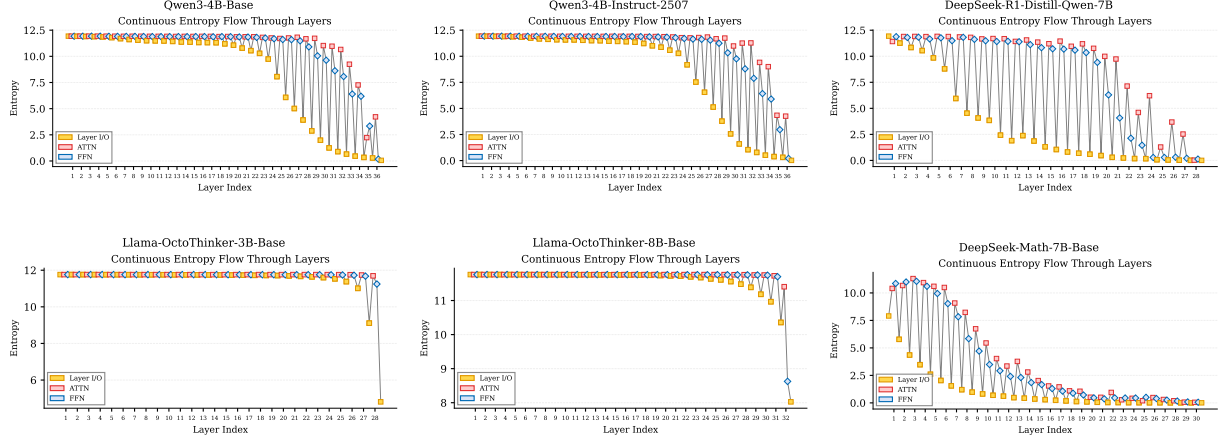


Figure 13: Continuous entropy dynamics of internal policy for additional models. The information flows from $\mathbf{H}^{l-1}$ into $\mathbf{A}^l$, $\mathbf{F}^l$, then to the next layer $\mathbf{H}^l$.

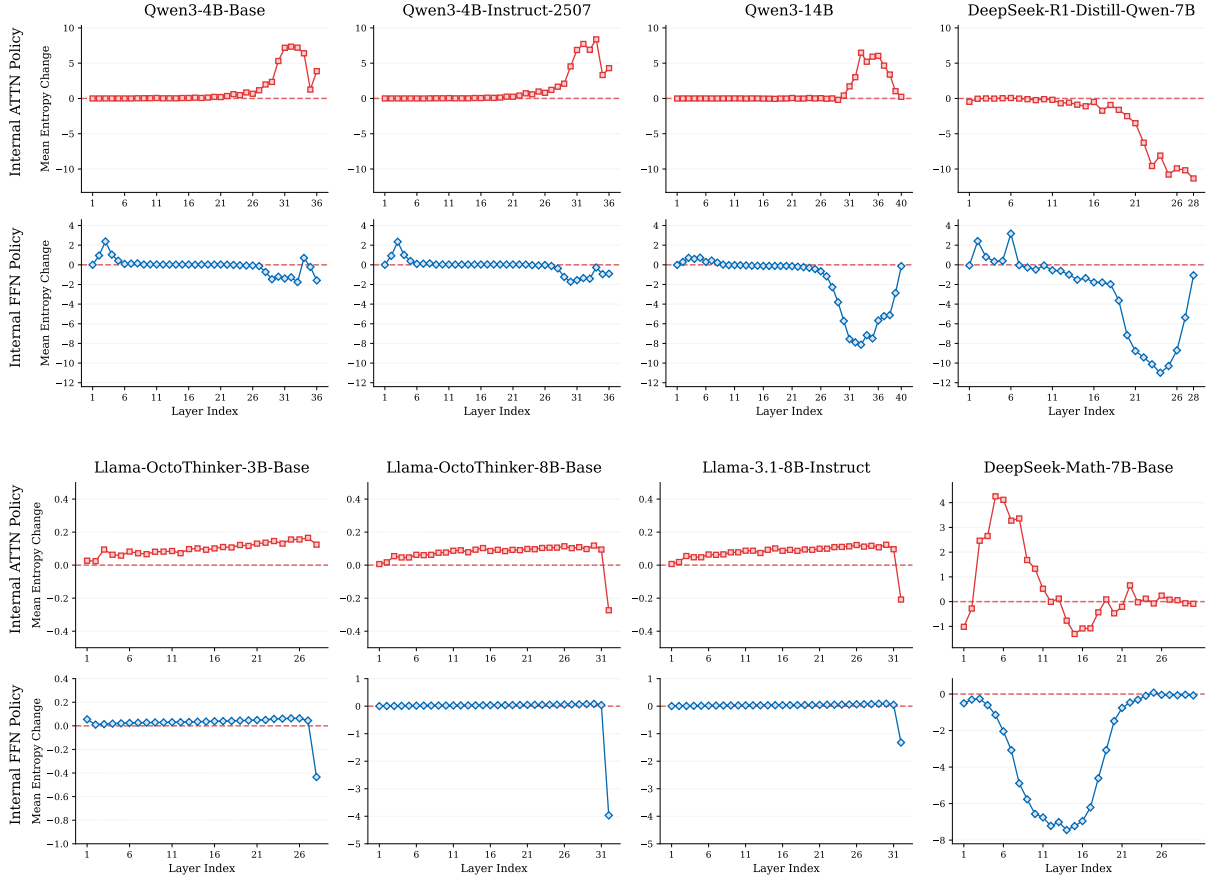


Figure 14: Entropy change dynamics of internal policy with more models.

B.6 Pass@K Performance across Datasets

We further provide a detailed version of Figure 6 across AMC23, MATH500, AIME24, and AIME25. Our proposed BuPO consistently outperforms the vanilla GRPO baseline.

C Use of AI Assistants

AI assistants were used only for minor language polishing and grammar refinement. All research ideas, technical contributions, experiments, analyses, and final claims were developed and verified by the authors.

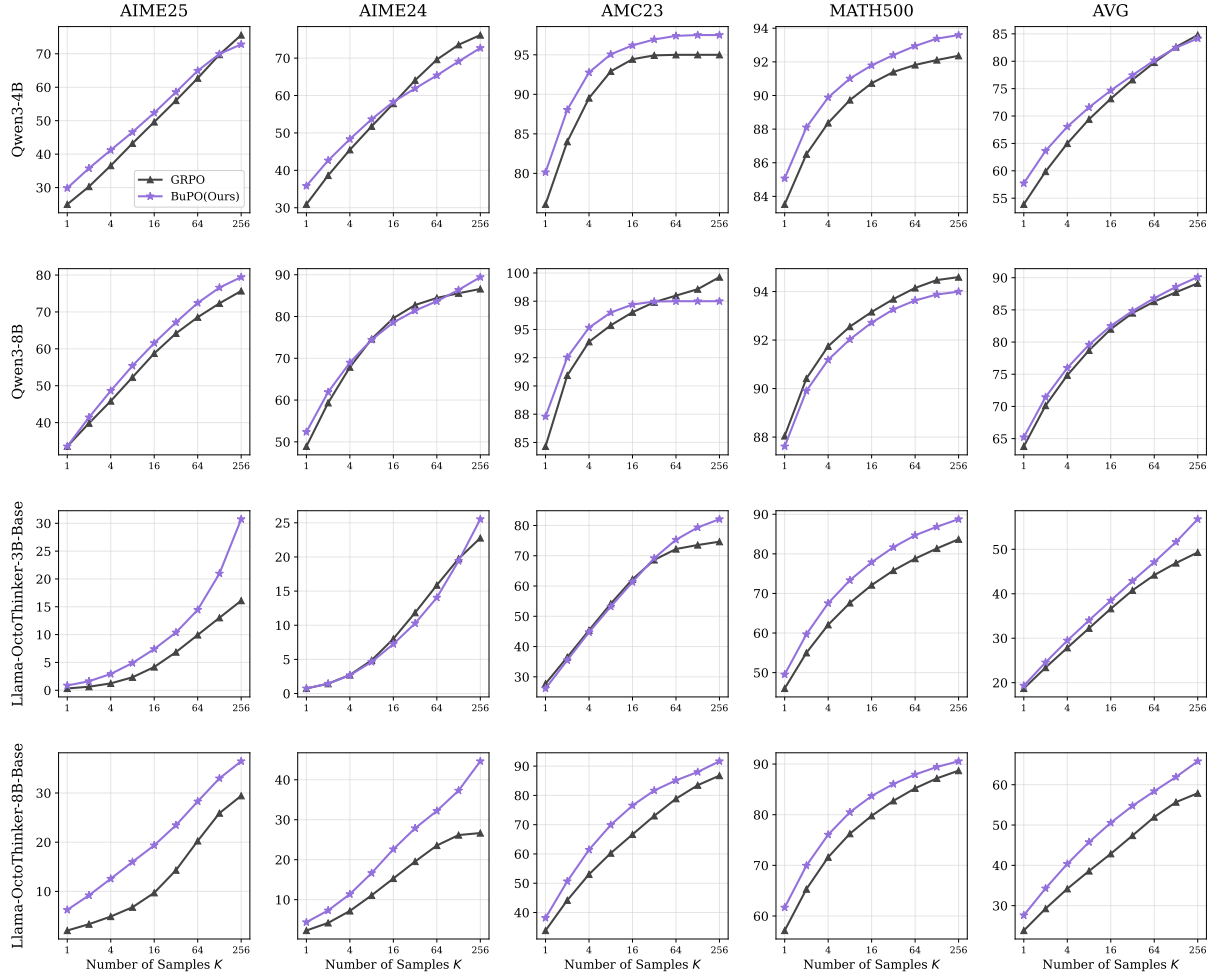


Figure 15: Pass@K results on MATH500, AMC23, AIME24 and AIME25. To reduce evaluation variance, we set $n = 300$.